NFTR: FROM PROVABLE MODE-AVERAGING TO GEODESIC SUBGOAL SELECTION IN OFFLINE GOAL-CONDITIONED RL

Anonymous authors
Paper under double-blind review

ABSTRACT

Hierarchical Implicit Q-Learning (HIQL), an offline goal-conditioned RL method, selects subgoals by value-function advantages alone. This rule has two coupled failure modes. *Optimistic bias* treats lucky stochastic outcomes as skillful choices, and *mode collapse* reduces a multi-modal subgoal distribution to a single Gaussian mean that often falls in unreachable regions. We propose **NFTR** (**N**ormalizing **F**low subgoal policies with **T**riangle-slack **R**eweighting). A conditional Normalizing Flow replaces the Gaussian policy. A closed-form mode-averaging result identifies the Gaussian limitation, while conditional flows support multi-modal weighted maximum likelihood and direct sampling. A triangle slack score, computed from a jointly trained MRN distance head whose triangle inequality is guaranteed architecturally, corrects the AWR weight multiplicatively and measures a detour inside the learned geometry without requiring exact distance recovery. Triangle-slack vanishes on geodesics in deterministic MDPs and remains a conservative upper bound on composability violation under stochastic dynamics. The RWDR objective preserves AWR's population-level monotonic improvement and admits a three-term suboptimality decomposition. On OGBench the flow carries the larger share of the gain, while the full co-trained configuration adds task-dependent gains. The combined method avoids Gaussian mode averaging and performs well on stochastic tasks.

1 INTRODUCTION

Offline goal-conditioned reinforcement learning (GCRL) (Park et al., 2025a) aims to learn policies that can reach arbitrary goals from static datasets without environment interaction. This setting is particularly valuable for robotics and autonomous systems where online data collection is expensive or dangerous. Among recent advances, Hierarchical Implicit Q-Learning (HIQL) (Park et al., 2023) has achieved state-of-the-art performance by decomposing goal-reaching into high-level subgoal selection and low-level action execution, using a shared value function to guide both levels.

Two coupled failures of HIQL motivate this work. First, because the high-level subgoal is selected purely by the learned value function, stochastic environments produce **optimistic bias**. A trajectory reaching the goal through favorable randomness is treated identically to one succeeding through repeatable reachability, so off-path subgoals receive inflated AWR weights. Second, because the high-level policy is a unimodal Gaussian, multi-modal subgoal distributions induced by branching corridors or stochastic outcomes trigger **mode collapse**. The Gaussian mean is regressed toward a convex combination of mode centers, which often falls inside an unreachable region rather than on any valid path. Together these failures raise a single question. *Can we learn subgoals that are simultaneously high-value, reliably reachable, and faithful to multi-modal path structure?*

We answer this question with **NFTR** (**N**ormalizing **F**low subgoal policies with **T**riangle-slack **R**eweighting). A conditional Normalizing Flow replaces HIQL's Gaussian high-level policy, and a triangle-slack score derived from an architecturally inequality-preserving quasimetric reweights the AWR objective, so geometric inconsistency acts as a mechanistic proxy for unreliable lucky transitions. We order our contributions by empirical weight. A closed-form diagnosis makes an expressive high-level class a consequence of the objective. A placement result puts the flow at the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

high level alone. Triangle-slack reweighting acts on the AWR weight, not on the value function. Two concurrent methods also insert flows into HIQL at both levels, and Appendix F.1 compares against them. On OGBench (Park et al., 2025a), large gains occur on teleport navigation and PointMaze medium stitching with multi-modal subgoals.

2 PRELIMINARIES

2.1 OFFLINE GOAL-CONDITIONED REINFORCEMENT LEARNING

We consider a controlled Markov process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \gamma, \mu)$ with state space $\mathcal{S}$, action space $\mathcal{A}$, transition dynamics $P(s'|s,a)$, discount factor $\gamma \in (0,1)$, and initial state distribution $\mu(s_0)$. A goal-conditioned policy $\pi(a|s,g)$ aims to reach goal state $g \in \mathcal{S}$ from current state s. The goal-conditioned value function $V^\pi(s,g)$ is the expected discount accumulated until g is first reached from s under policy π:

$$V^\pi(s,g) = \mathbb{E}_\pi\left[\gamma^{T_g} \mid s_0 = s\right], \qquad T_g = \inf\{t \geq 0 : s_t = g\}, \tag{1}$$

with $\gamma^\infty = 0$. This gives $V^\pi(g,g) = 1$, so $d^*(s,g) = -\log V^*(s,g)$ satisfies $d^*(g,g) = 0$ as Definition 1 requires. In offline GCRL, we have access only to a static dataset $\mathcal{D} = \{(s_t, a_t, s_{t+1})\}_{t=1}^N$ collected by an unknown behavior policy β, with goals typically sampled from future states via hindsight relabeling (Andrychowicz et al., 2017). The challenge is to learn a policy π that maximizes goal-reaching performance without additional environment interaction, requiring careful handling of distribution shift between the learned policy and the behavior policy (Levine et al., 2020).

2.2 HIQL: HIERARCHICAL IMPLICIT Q-LEARNING

HIQL (Park et al., 2023) factorizes goal-reaching into a high-level policy $\pi^H(z \mid s, g)$ that proposes the latent code $z = \phi([s; w])$ of a k-step waypoint w, and a low-level policy $\pi^L(a \mid s, z)$ that drives the agent toward it. The encoder $\phi : \mathcal{S} \times \mathcal{S} \to \mathbb{R}^d$ is an intermediate layer of the parameterized value function $V(s, \phi([s; g]))$, learned end-to-end with no separate representation objective; HIQL's Proposition 5.1 shows this representation is sufficient for optimal control in deterministic MDPs.

The shared value function is fitted with the action-free expectile-regression variant of IQL (Kostrikov et al., 2022; Park et al., 2023):

$$\mathcal{L}_V \;=\; \mathbb{E}_{(s,s')\sim\mathcal{D},\, g\sim p(g|\tau)}\left[L_\tau^2\big(r(s,g) + \gamma V^{\mathrm{tgt}}(s',g) - V_\theta(s,g)\big)\right], \tag{2}$$

with $L_\tau^2(u) = |\tau - \mathbf{1}(u < 0)| \cdot u^2$ for $\tau \in (0.5, 1)$. The high-level policy is then extracted by advantage-weighted regression (Peng et al., 2019):

$$\mathcal{L}_{\pi^H} \;=\; -\,\mathbb{E}_{(s,w,g)\sim\mathcal{D}}\left[\exp\big(\alpha_H A_{\mathrm{env}}(s,w,g)\big) \cdot \log \pi^H\big(\phi([s;w]) \;\big|\; s, g\big)\right], \tag{3}$$

where $A_{\mathrm{env}}(s,w,g) = V_\theta(w,g) - V_\theta(s,g)$ and $\alpha_H > 0$. Following HIQL's Section 5.2, the regression target is the latent $z = \phi([s;w])$ rather than raw w, a choice required for image-based and consistently helpful for state-based settings; NFTR keeps it unchanged.

Two technical hooks for our analysis. HIQL's hierarchy is justified by a signal-to-noise argument (Park et al., 2023, Proposition 4.1) that is silent on the high-level policy class, leaving the unimodal Gaussian inherited from POR (Xu et al., 2022) unanalyzed. The action-free objective in Equation (2) is also unbiased only under deterministic dynamics (Park et al., 2023, Section 3), and AWR in Equation (3) exponentially amplifies any resulting value overestimation. Sections 3.1 and 3.2 exploit these two openings.

2.3 TRIANGLE INEQUALITY AS A STRUCTURAL PRIMITIVE

We single out one structural property, the triangle inequality, from the broader machinery of quasimetric learning. Earlier work (Wang et al., 2023; Liu et al., 2023; Myers et al., 2024; 2025b) treats the quasimetric as an *estimation target*, where downstream performance scales with how accurately the distance is recovered. We instead treat the inequality alone as a *structural prior* on the AWR weight, regardless of whether the underlying distance is recovered accurately.

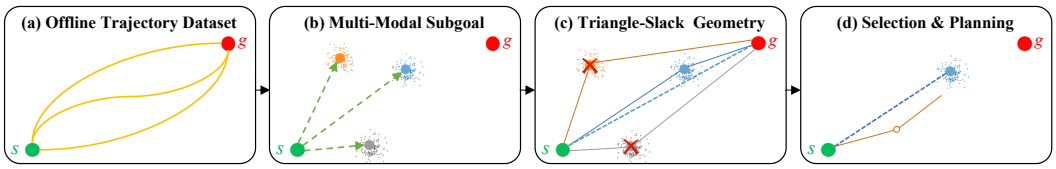


Figure 1: **Overview of NFTR.** (a) Offline trajectories. (b) A Normalizing Flow high-level policy models a multi-modal subgoal distribution. (c) A quasimetric d_θ induces the slack $\Delta(s, w, g) = \mathrm{ReLU}(d(s, w)+d(w, g)-d(s, g))$, which downweights non-composable subgoals. (d) Execution yields a feasible stitched path. RWDR combines A_{env} from V_θ and Δ from d_θ to train the policy.

Definition 1 (Quasimetric)**.** A function $d : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ is a **quasimetric** if it satisfies: (1) $d(s, s) = 0$, (2) $d(s, g) > 0$ for $s \neq g$, and (3) $d(s, g) \leq d(s, w) + d(w, g)$ for all s, w, g. Unlike metrics, quasimetrics need not be symmetric.

In what follows, d_θ stands for a distance whose triangle inequality is guaranteed at the architectural level. The concrete MRN parameterization and its training losses are deferred to Sections 3.4 and 4. Two consequences of this construction matter for our framework. First, the inequality holds *architecturally*, so the slack is non-negative before any training. Second, distance training shapes a *geometry inside which detours can be measured*, without requiring exact recovery of reachability distances. The first property is exact, and Appendix F.5 measures the second empirically.

2.4 Normalizing Flows (NFs)

NFs (Dinh et al., 2014; 2017; Papamakarios et al., 2021; Ghugare & Eysenbach, 2025) define probability distributions by transforming simple base distributions through invertible functions, enabling both exact density computation and efficient sampling.

Definition 2 (Normalizing Flow)**.** A Normalizing Flow defines a distribution $p_\theta(z)$ over $z \in \mathbb{R}^d$ through an invertible transformation $f_\theta : \mathbb{R}^d \to \mathbb{R}^d$ applied to a base distribution $p_{\text{base}}(\epsilon)$:

$$z = f_\theta^{-1}(\epsilon), \quad \epsilon \sim p_{\text{base}}(\epsilon). \tag{4}$$

The density is computed via the change-of-variables formula:

$$\log p_\theta(z) = \log p_{\text{base}}(f_\theta(z)) + \log \left| \det \frac{\partial f_\theta}{\partial z} \right|. \tag{5}$$

A conditional flow conditions the transformation f_θ on context c to model $p_\theta(z \mid c)$, which we instantiate with RealNVP (Dinh et al., 2017) affine coupling layers in Section 3.4.

3 Method: NFTR

We present NFTR, which addresses HIQL's limitations through two complementary innovations: a Normalizing Flow high-level subgoal policy for multi-modal subgoal distributions (Section 3.1), and triangle-slack composability reweighting for filtering geometrically inconsistent subgoals (Section 3.2). We then provide theoretical analysis of the combined method (Section 3.3) and describe implementation details (Section 3.4). The overall architecture is illustrated in Figure 1.

3.1 Normalizing Flow High-Level Subgoal Policy

Two facts about HIQL's high-level training set drive the analysis. The latent code $z = \phi([s; w])$ is the deterministic image of the k-step waypoint $w = s_{t+k}$ (Park et al., 2023, Algorithm 1), and for a fixed (s, g) appearing in many trajectories the empirical distribution of w is shaped by trajectory-level path multiplicity rather than single-step stochasticity. Branching corridors, teleporters that randomize landing sites, and observation noise that spreads the same physical waypoint across ϕ-space all produce disjoint waypoint clusters separated by unreachable regions $\mathcal{S}_{\text{wall}} \subset \mathbb{R}^d$. We model this as a state-dependent mixture

$$p_{\mathcal{D}}(z \mid s, g) \;=\; \sum_{m=1}^{M(s,g)} \pi_m(s, g)\, q_m(z \mid s, g), \tag{6}$$

with path-connected, pairwise disjoint component supports $\mathrm{supp}(q_m)$. We treat $\mathcal{S}_{\text{wall}}$ as the empirical ϕ-image of the physical unreachable region observed in our visualizations (Figure 7), since HIQL's representation-sufficiency (Park et al., 2023, Proposition 5.1) is only exact in the deterministic case.

The next theorem characterizes the regime in which HIQL's POR-inherited unimodal Gaussian fails. To our knowledge it is the first closed-form characterization of HIQL's mode-collapse limitation.

Theorem 3.1 (Mode-Averaging of Gaussian High-Level Subgoal Policies)**.** *Let $\pi_\theta^H(z \mid s, g) = \mathcal{N}(z;\, \mu_\theta(s,g),\, \sigma^2 I_d)$ with $\sigma > 0$ fixed and μ_θ unconstrained, trained by HIQL's AWR objective*

$$\mathcal{L}^{\text{AWR}}(\theta) \;=\; -\,\mathbb{E}_{(s,w,g)\sim\mathcal{D}}\left[e^{\alpha_H A^H(s,w,g)}\,\log\mathcal{N}\big(\phi([s;w]);\,\mu_\theta(s,g),\,\sigma^2 I_d\big)\right], \tag{7}$$

with $\alpha_H > 0$ and $e^{\alpha_H A^H}$ bounded above and integrable under $p_{\mathcal{D}}$ (enforced by the weight clipping in Algorithm 1). Let Equation (6) hold with $M(s,g) \geq 2$ and write $w(z)$ for the deterministic preimage of $z = \phi([s;w])$. Define $p^\star_{s,g}(z) := Z^{-1}_{s,g} e^{\alpha_H A^H(s,w(z),g)} p_{\mathcal{D}}(z \mid s,g)$ with normalizer $Z_{s,g} = \mathbb{E}_{p_{\mathcal{D}}}[e^{\alpha_H A^H} \mid s,g]$. The unique minimizer of Equation (7) in $\mu_\theta(s,g)$ is

$$\mu^\star_\theta(s,g) \;=\; \mathbb{E}_{z\sim p^\star_{s,g}}[z] \;=\; \sum_{m=1}^{M(s,g)} \tilde{\pi}_m(s,g)\,\tilde{\mu}_m(s,g), \tag{8}$$

where $\tilde{\mu}_m$ and $\tilde{\pi}_m$ are the AWR-tilted modal centers and weights (full forms in Appendix D.1). Hence $\mu^\star_\theta(s,g) \in \mathrm{conv}\{\tilde{\mu}_m(s,g)\}$. When the tilted weights produce $\mu^\star_\theta(s,g) \in \mathcal{S}_{\text{wall}}$, generic for approximately balanced modes separated by the wall, the Gaussian places mass on $\mathcal{S}_{\text{wall}}$ bounded below by a strictly positive constant for every σ in any fixed compact range.

Theorem 3.1 states that the AWR-optimal Gaussian mean is a fixed convex combination of AWR-tilted modal centers, falls inside any region those centers straddle, and irreducibly leaks mass into that region; Figure 2 illustrates this failure on a two-corridor case alongside the corresponding NF behavior. The full proof is in Appendix D.1.

Remark 3.2 (The diagnosis does not depend on the density). Equation (7) is a weighted maximum likelihood objective, but its minimizer in μ_θ is unchanged when the Gaussian head is fitted by weighted least-squares regression. Step 1 of Appendix D.1 reduces Equation (7) to $\frac{1}{2\sigma^2}\,\mathbb{E}[e^{\alpha_H A^H}\|z-\mu_\theta\|_2^2]$ up to a term free of μ_θ. The convex combination in Equation (8) and the wall-mass conclusion therefore hold under either route, so the scoping below concerns the admissible model class rather than the failure described here.

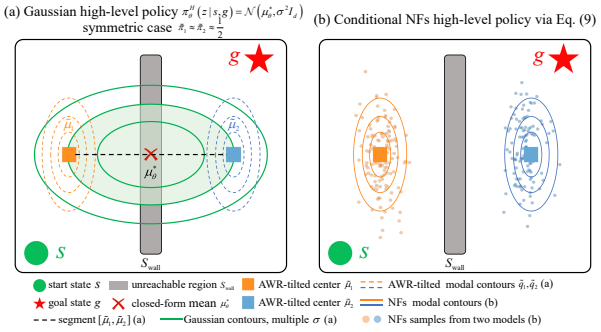


Figure 2: **Geometric illustration of Theorem 3.1. (a)** In the symmetric two-mode case $\widetilde{\pi}_1 \approx \widetilde{\pi}_2 \approx \frac{1}{2}$, the AWR-optimal Gaussian mean μ^*_θ from Equation (8) sits at the midpoint of $[\widetilde{\mu}_1, \widetilde{\mu}_2]$ inside $\mathcal{S}_{\text{wall}}$, and the Gaussian (green ellipses at multiple σ) leaks irreducible mass onto $\mathcal{S}_{\text{wall}}$. **(b)** A conditional NF (Equation (9)) places mass on both corridors and drives the wall mass arbitrarily small as capacity grows (Papamakarios et al., 2021).

Policy choice. The mode-averaging result motivates a policy that represents multiple subgoal modes (R1). Weighted maximum likelihood calls for an exact log density (R2). We also prefer direct sampling at execution time (R3). A conditional NF meets these requirements:

$$\log \pi^H_\theta(z \mid s,g) \;=\; \log\mathcal{N}\big(f_\theta(z;\,s,g);\,0,I_d\big) \;+\; \log\left|\det\frac{\partial f_\theta(z;\,s,g)}{\partial z}\right| \tag{9}$$

This parameterization provides an exact density (R2), represents multiple modes (R1) (Papamakarios et al., 2021), and permits direct sampling through f_θ^{-1} (R3). The NF is therefore a practical fit for this objective. Weighted regression can train other expressive generators without exact density evaluation. Appendices F.1 and F.3 examines those alternatives. Figure 2 illustrates the two-corridor case, and Figure 7 (Appendix F) shows learned policy densities on `antmaze-teleport-navigate`. Flow Q-Learning (FQL) (Park et al., 2025b) and NF-RLBC (Ghugare & Eysenbach, 2025) parameterize action policies rather than high-level subgoals.

3.2 TRIANGLE-SLACK COMPOSABILITY REWEIGHTING

Multi-modal modeling alone does not address optimistic bias. The AWR advantage $A_{\text{env}}(s, w, g) = V(w, g) - V(s, g)$ tells us whether w *appears* closer to g under the learned value function, but it cannot separate two qualitatively different events that both produce a positive advantage in the dataset. The first is a subgoal reached along a reliable path. The second is a subgoal reached via a lucky transition (a teleporter outcome, an observation-noise realization). Both inflate A_{env}, and AWR amplifies them identically through $\exp(\alpha_H A_{\text{env}})$. The root cause is that V answers a sample-level question, namely "did this subgoal lead to success in this data?" rather than the population-level question we actually care about, namely "is this subgoal reliably on the typical path from s to g?"

Triangle-slack as a geometric consistency signal. We introduce triangle-slack as a consistency signal that identifies composable subgoals independently of value estimates. Given a learned quasimetric distance d_θ (detailed in Section 3.4), we define the triangle-slack of subgoal w for state-goal pair (s, g) as

$$\Delta(s, w, g) \;=\; \max\big(0,\; d_\theta(s, w) + d_\theta(w, g) - d_\theta(s, g)\big). \tag{10}$$

By the triangle inequality property of quasimetrics, $\Delta \geq 0$ always holds, with $\Delta \approx 0$ when w lies on a geodesic path from s to g and $\Delta > 0$ when reaching g via w requires a detour compared to the direct path. Figure 3 illustrates this geometry: composable waypoints incur near-zero slack while non-composable ones produce positive slack indicating a detour.

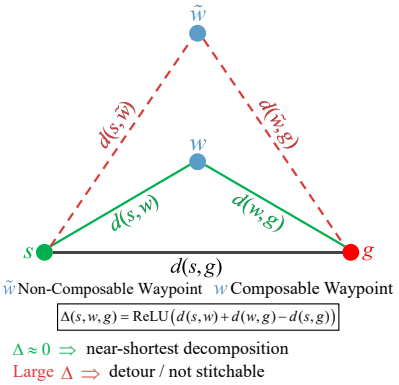


Figure 3: **Triangle-slack geometry.** A composable waypoint w approximately satisfies the triangle equality, i.e., $d(s, w) + d(w, g) \approx d(s, g)$, yielding a small triangle slack $\Delta(s, w, g) \approx 0$ (green edges). In contrast, a non-composable waypoint $\tilde{w}$ violates this relation, producing a large slack $\Delta(s, \tilde{w}, g) > 0$ (red dashed edges), which indicates a detour / poor composability and should be downweighted by RWDR.

We now make these geometric properties precise. We first formalize the notion of a geodesic subgoal:

Definition 3 (Geodesic Subgoal)**.** A subgoal w is **geodesic** for (s, g) under quasimetric d if $d(s, g) = d(s, w) + d(w, g)$.

The following proposition characterizes the four key properties of triangle-slack as a composability signal.

Proposition 3.3 (Slack Characterization)**.** *For any quasimetric d satisfying Definition 1 and any subgoal $w \in \mathcal{S}$:*

1. $\Delta(s, w, g) \geq 0$ *for all* $s, w, g \in \mathcal{S}$.
2. $\Delta(s, w, g) = 0$ *if and only if* w *is geodesic for* (s, g).
3. $\Delta(s, w, g) > 0$ *if and only if reaching* g *via* w *requires a strict detour, i.e.,* $d(s, w) + d(w, g) > d(s, g)$.
4. *In a deterministic MDP, under the optimal distance* $d^*(s, g) = -\log V^*(s, g)$*, geodesic subgoals are exactly those on optimal paths. In a general (stochastic) MDP, the multiplicative value decomposition relaxes to* $V^*(s, g) \geq V^*(s, w) \cdot V^*(w, g)$*, so* $d^*(s, g) \leq d^*(s, w) + d^*(w, g)$ *becomes a (strict) inequality for off-path* w*, and triangle-slack acts as a* conservative upper bound *on composability violation.*

The proof is in Appendix D.2. Parts (1) to (3) say triangle-slack vanishes exactly on geodesics, and part (4) says it remains a conservative upper bound on composability violation under stochastic dynamics. The link to lucky-transition filtering rests on two properties. Non-negativity is architectural, so the correction can only remove weight and never reward a shortcut. A larger detour inside the learned geometry of d_θ gives a larger slack, and Appendix F.5 measures its association with unreliable transitions empirically.

The RWDR objective. Building on Equation (10), we combine the value advantage and triangle-slack penalty into the RWDR weighting scheme:

$$\log w(s, w, g) \;=\; \alpha_H \cdot A_{\text{env}}(s, w, g) - \kappa \cdot \Delta(s, w, g), \tag{11}$$

where $\kappa \geq 0$ controls the geometric penalty. Setting $\kappa = 0$ recovers the value-only weight. For $\kappa > 0$, the weight $w(s, w, g)$ is high when the subgoal has both high value advantage (making progress toward the goal) and low triangle-slack (geometrically consistent with being on-path). When the two signals conflict, RWDR can downweight a high-value subgoal with large slack. The final high-level policy loss becomes:

$$\mathcal{L}_{\pi^H} = -\mathbb{E}_{(s,w,g)\sim\mathcal{D}}\left[w(s, w, g) \cdot \log \pi^H_{\text{NF}}(\phi([s; w]) \mid s, g)\right], \tag{12}$$

where we apply stop-gradient to both V_θ and d_θ when computing weights, treating them as fixed scorers for each policy update. This prevents gradients from the policy loss from updating the value and distance heads. It does not freeze those heads or the value encoder ϕ across training iterations; ϕ continues to evolve under the value loss, so the NF regression targets can drift during training.

The reweighting in Equation (11) can be equivalently expressed as a multiplicative correction to the AWR weights, which we record as a corollary for later use:

Assumption 3.4 (Geometric Consistency)**.** *The distance d_θ satisfies the triangle inequality exactly:*

$$d_\theta(s, g) \leq d_\theta(s, w) + d_\theta(w, g), \quad \forall s, w, g \in \mathcal{S}. \tag{13}$$

This property is guaranteed by the MRN architecture and holds regardless of training.

Corollary 3.5 (Geometric Regularization Effect)**.** *Under Assumption 3.4, for any distance function d_θ satisfying the triangle inequality and $\kappa \geq 0$, the raw RWDR weight is no greater than the value-only weight:*

$$w_{\textit{RWDR}}(s, w, g) = w_{\textit{HIQL}}(s, w, g) \cdot \exp\big(-\kappa\,\Delta(s, w, g)\big). \tag{14}$$

If $\kappa > 0$ and $\Delta(s, w, g) > 0$, then $w_{\textit{RWDR}}(s, w, g) < w_{\textit{HIQL}}(s, w, g)$ before clipping and batch normalization.

The proof is provided in Appendix D.3. For $\kappa > 0$, RWDR exponentially downweights positive-slack subgoals in the raw weight. The local properties established here describe the geometric correction. The system-level consequences of combining the NF high-level policy from Section 3.1 with the RWDR objective in Equation (12) are analyzed in the next section.

3.3 Theoretical Analysis of NFTR

Sections 3.1 and 3.2 established two component-level results: NF parameterization avoids mode-averaging at the high-level policy, and triangle-slack RWDR functions as a geometric composability filter (Proposition 3.3 and Corollary 3.5). We now combine both into system-level guarantees, under two regularity assumptions.

Assumption 3.6 (Single-Policy Concentrability)**.** *There exists $C_\pi < \infty$ such that $d^{\pi^{H,*}}(s, w)/d^\beta(s, w) \leq C_\pi$ for all (s, w) in the support of the optimal high-level policy $\pi^{H,*}$, where d^π is the state-subgoal visitation under π and d^β is the dataset distribution.*

Assumption 3.7 (Value Estimation Quality)**.** *For high-level samples $(s, w, g) \sim \mathcal{D}$, the learned value satisfies $\mathbb{E}[(V_\theta(s, g) - V^*(s, g))^2] \leq \epsilon_V$ and $\mathbb{E}[(V_\theta(w, g) - V^*(w, g))^2] \leq \epsilon_V$.*

Combined with Assumption 3.4, these yield the following decomposition.

Theorem 3.8 (RWDR Suboptimality Decomposition)**.** *Assume $|\hat{A}_{\text{env}}|, |A^*_{\text{env}}| \leq A_{\max}$ and $d_\theta \leq D_{\max}$ (so $\Delta \leq 2D_{\max}$). Under Assumptions 3.4, 3.6 and 3.7, the population minimizer $\hat{\pi}^H$ of the empirical RWDR objective (Equation (12)) with n samples satisfies, with high probability,*

$$J(\pi^{H,*}) - J(\hat{\pi}^H) \leq \underbrace{\mathcal{O}(\sqrt{\epsilon_V})}_{\textit{value error}} + \underbrace{\mathcal{O}\big(\sqrt{C_\pi\,\mathcal{C}(\Pi)/n}\big)}_{\textit{sampling error}} + \underbrace{R_\kappa(d_\theta)}_{\textit{regularization}}, \tag{15}$$

where $R_\kappa(d_\theta) := \kappa A_{\max}\,\mathbb{E}_{(s,w,g)\sim\pi^{H,*}}[\Delta(s, w, g)] \leq 2\kappa A_{\max} D_{\max}$ is the residual bias of the geometric reweighting on optimal subgoals, and $\mathcal{C}(\Pi)$ is the policy-class complexity ($\log|\Pi|$ for finite Π, replaced by Rademacher or covering-number bounds for parameterizations such as RealNVP). The hidden constant in the value-error term may depend on the fixed α_H, κ, $A_{\max}$ and $D_{\max}$.

The proof is in Appendix D.4. The three terms are IQL value-estimation error, the standard $\mathcal{O}(1/\sqrt{n})$ sampling error, and a geometric regularization that vanishes when d_θ aligns with optimal distances (since $\Delta = 0$ on geodesics by Proposition 3.3 (2)) and is otherwise uniformly bounded. Together with Theorem 3.1, this characterizes how NFTR's two components combine: the NF makes $\hat{\pi}^H$ realizable, while RWDR controls which subgoals dominate the minimization via $R_\kappa(d_\theta)$.

Proposition 3.9 (Population-Level Monotonic Improvement)**.** *Assuming exact weighted-MLE optimization over* Π*, the RWDR objective inherits AWR's monotonic improvement: for any* $\pi^H_{\text{old}} \in \Pi$ *and the exact weighted-MLE solution* π^H_{new}*,*

$$\text{KL}(\pi^{H,*}_{\mathcal{D}} \,\|\, \pi^H_{\text{new}}) \;\le\; \text{KL}(\pi^{H,*}_{\mathcal{D}} \,\|\, \pi^H_{\text{old}}), \tag{16}$$

where $\pi^{H,*}_{\mathcal{D}}$ *is the RWDR-weighted in-distribution optimum.*

The proof is in Appendix D.5. Like all AWR-based methods (HIQL, SAW, OTA), Proposition 3.9 does not transfer verbatim to SGD on a non-convex RealNVP parameterization, a gap that originates from π^H rather than from RWDR; the evaluation success curves in Appendix F.10 illustrate practical stability but do not establish per-step loss or return improvement.

3.4 NFTR Algorithm

NFTR alternates per step an IQL value update, an MRN distance update, and a weighted-MLE update of π^H_{NF} together with the low-level AWR actor. The high-level update uses the RWDR weight from Equation (11), combining A_{env} with the triangle slack Δ from the architecturally inequality-preserving d_θ. We parameterize d_θ with an MRN (Liu et al., 2023) trained by a contrastive plus Bellman-consistency loss, and π^H_{NF} with $K{=}4$ RealNVP coupling layers. The total objective is

$$\mathcal{L} \;=\; \mathcal{L}_V + \mathcal{L}_{\pi^L} + \mathcal{L}_{\pi^H} + \lambda_d \mathcal{L}_d, \tag{17}$$

with $\mathcal{L}_V$ the IQL expectile loss (Equation (2)), $\mathcal{L}_{\pi^L}$ the low-level AWR loss, $\mathcal{L}_{\pi^H}$ the RWDR-weighted NF loss (Equation (12)), and $\mathcal{L}_d$ the auxiliary MRN distance loss. The full NFTR configuration (A0) sets $\lambda_d = 1$ and jointly trains the MRN distance head with the contrastive and Bellman-consistency losses. When computing each high-level policy weight, we stop gradients through d_θ. The head continues to receive updates through $\mathcal{L}_d$. The frozen-head control (A2) instead sets $\lambda_d = 0$ and keeps d_θ at its initialization. RWDR adds distance-head evaluations, a slack cap $\Delta_{\max}$ and a scalar coefficient κ, while weight clipping and per-batch normalization are inherited AWR practice. We select κ per environment on a held-out split. Appendix F.8 reports the sweep and identifies the settings matching the reported NFTR results. Full pseudocode and hyperparameters are in Appendices C and E.

4 Experiments

We evaluate NFTR on OGBench (Park et al., 2025a), focusing on three scenarios where HIQL's limitations are most pronounced: (1) stochastic environments with random transitions, (2) stitching tasks requiring trajectory composition, and (3) manipulation tasks with complex dynamics.

4.1 Experimental Setup

We evaluate on OGBench's locomotion (PointMaze, AntMaze) and manipulation (Cube, Scene) suites, covering the task categories `navigate` (standard goal-reaching), `stitch` (short trajectory segments that must be composed), `teleport` (stochastic dynamics with random teleportation), `explore` (low-quality exploratory data), and `play`/`noisy` (manipulation under varying observation noise). We compare against hierarchical methods HIQL (Park et al., 2023), SAW (Zhou & Kao, 2025), OTA (Ahn et al., 2025), flat value- and distance-based methods QRL (Wang et al., 2023), CRL (Eysenbach et al., 2022), TMD (Myers et al., 2025b), MQE (Zheng et al., 2026), flow-based methods GCFQL and H-GCFQL (Park et al., 2025b), and NF-based methods GC-NF-RLBC and H-GC-NF-RLBC (Ghugare & Eysenbach, 2025). Two further comparisons appear in Appendices F.1 and F.2. Our runs report success rates (%) over 5 test-time goals and 8 seeds, with standard deviations. HIQL results are quoted from the original paper.

4.2 Main Results

Stochastic, stitching, and manipulation. Tables 1 and 2 cover NFTR's three target regimes. On PointMaze and AntMaze teleport tasks, triangle-slack complements value estimates: the MRN reflects typical reachability rather than lucky outcomes, so unreliable subgoals can incur higher

Table 1: **OGBench Locomotion Results.** Success rate (%) on stochastic and stitching tasks, mean±std. Our evaluations use 8 seeds. NFTR has the highest or second-highest mean on all tasks. No per-task statistical significance is inferred from these summary statistics. Per row, the highest mean is **orangebold** and the second-highest is *underlined*.

Task	NFTR	HIQL	SAW	OTA	QRL	CRL	TMD	MQE	GCFQL	H-GCFQL
PointMaze (Stochastic & Stitching)										
pointmaze-teleport-navigate	**53.8 ± 5.9**	18 ± 4	37.5 ± 4.5	41.2 ± 6.9	4 ± 4	24 ± 6	22.7 ± 3.5	24.3 ± 3.5	*44.1 ± 4.2*	39.8 ± 1.7
pointmaze-teleport-stitch	**41.7 ± 4.5**	34 ± 4	*39.1 ± 6.8*	32.2 ± 3.9	9 ± 5	4 ± 3	29.3 ± 2.2	38.5 ± 5.6	**23.2 ± 1.4**	36.3 ± 3.2
pointmaze-medium-stitch	**98.1 ± 1.2**	74 ± 6	**87.1 ± 2.2**	75 ± 5	80 ± 12	0 ± 1	0.1 ± 0.1	67.6 ± 3.2	1.8 ± 0.2	49.9 ± 3.1
AntMaze (Stochastic & Stitching)										
antmaze-teleport-navigate	*52.0 ± 0.6*	42 ± 3	41.3 ± 1.4	46.2 ± 0.8	35 ± 5	**53 ± 2**	25.4 ± 2.9	46.6 ± 3.3	22.1 ± 8.7	39.8 ± 1.9
antmaze-teleport-stitch	**44.0 ± 5.5**	36 ± 2	33.8 ± 0.8	32.9 ± 3.3	24 ± 5	31 ± 4	10.7 ± 1.8	*37.4 ± 3.2*	14.7 ± 2.6	16.9 ± 2.9
visual-antmaze-teleport-navigate-v0	*46.4 ± 3.2*	37 ± 2	42.7 ± 2.3	37.1 ± 3.5	6 ± 3	**48 ± 2**	4.8 ± 0.8	5.8 ± 1.0	6.2 ± 3.1	4.6 ± 2.9
visual-antmaze-teleport-stitch-v0	**38.9 ± 1.3**	*37 ± 4*	33.1 ± 1.3	20.0 ± 3.9	1 ± 2	32 ± 6	0.9 ± 0.4	0.9 ± 0.4	1.0 ± 0.8	1.0 ± 0.8

Table 2: **OGBench Manipulation Results.** Success rate (%) on cube and scene tasks. NFTR consistently improves over HIQL, with particularly large gains on noisy datasets. Per row, the best result is **orangebold** and the second-best is *underlined*.

Task	NFTR	HIQL	OTA	GC-NF-RLBC	H-GC-NF-RLBC
cube-single-play	**47.1 ± 9.9**	15 ± 3	12.8 ± 1.4	*18.7 ± 0.3*	9.1 ± 2.5
cube-single-noisy	**67.4 ± 0.9**	*41 ± 6*	39.7 ± 2.1	14.1 ± 0.8	18.3 ± 1.6
scene-play	**45.6 ± 3.8**	*38 ± 3*	19.4 ± 4.4	10.2 ± 1.2	20.4 ± 2.6
scene-noisy	**32.6 ± 5.8**	*25 ± 4*	12.1 ± 1.0	4.1 ± 0.2	11.3 ± 0.7

Table 3: **Ablation on Hierarchical Policy.** Success rate (%) with full NFTR compared to replacing both levels with Normalizing Flows (NF2-NFTR). Per row, the best result is **orangebold**.

Task	HIQL	NFTR	NF2-NFTR
pointmaze-teleport-navigate	18 ± 4	**53.8 ± 5.9**	45.5 ± 10.5
pointmaze-teleport-stitch	34 ± 4	**40.8 ± 8.8**	38.5 ± 7.0
antmaze-teleport-navigate	42 ± 3	**52.0 ± 0.6**	49.4 ± 1.7
antmaze-teleport-stitch	36 ± 2	**44.0 ± 5.5**	40.2 ± 4.8

slack. On medium stitching, the NF models disconnected waypoint clusters near trajectory endpoints; teleport-stitch combines this multi-modality with stochastic transitions. On manipulation, the NF captures alternative grasp and approach strategies, while RWDR can downweight high-slack waypoints associated with observation noise. This geometric signal is clearer in PointMaze than in higher-dimensional AntMaze (Sections 3.2 and 3.3). Adding a low-level NF (NF2-NFTR, Table 3) underperforms NFTR by 2.6–8.3 pp, supporting high-level-only placement.

Component comparisons. On the four teleport columns of Table 4, B1 exceeds the ablation baseline B0 by 10.6, 8.9, 37.1, and 12.6 pp. A0 exceeds B1 by 2.2, 3.8, 5.0, and 2.4 pp. The corresponding Gaussian configurations have positive mean differences of 0.9 to 5.1 pp. The flow provides the larger increment. These are descriptive means over 8 seeds, not per-task significance tests.

Comparison with distance-based methods. TMD, QRL, and CRL excel where precise reachability or robust contrastive value estimation is the bottleneck, especially in high-dimensional flat control. NFTR is complementary. It uses distance as a geometric scorer inside a hierarchy, which helps most when multi-modal subgoals and stochastic transitions co-occur.

4.3 Ablation Studies

Table 4 compares the policy class, RWDR weight, and distance-head training on four teleport tasks and one stitching task. The comparisons below use means over 8 seeds.

Q1: Does the Normalizing Flow policy improve over a Gaussian policy?

A1: B1 replaces B0's Gaussian high-level policy with an NF while retaining the value-only weight. Its mean success is higher on all five tasks, by 8.9 to 37.1 pp. The largest gain is on `pm-tel-nav`, from 11.7 to 48.8. On `pm-med-stitch`, success rises from 70.0 to 97.3. These large PointMaze changes are consistent with the mode-averaging diagnosis in Theorem 3.1. B1 is our ablation arm, distinct from the published NF-HIQL method in Table 10.

Q2: Does the full A0 configuration improve over B1?

A2: A0 adds a co-trained distance head and the RWDR weight to B1. Its mean success is higher on all five tasks, by 2.2, 3.8, 5.0, 2.4, and 0.8 pp in table order. The largest increment is on `pm-tel-nav`. This comparison does not isolate the slack penalty because the arms also differ in distance-head training. Moreover, A0 uses $\kappa = 0$ on `pm-tel-stitch` and `pm-med-stitch` (Table 15), where the slack term does not change the raw weight.

Q3: How do the co-trained and frozen-head configurations compare?

A3: With the NF policy and RWDR configuration, co-trained A0 has higher mean success than frozen-head A2 on all five tasks. The differences are 2.0, 1.9, 3.9, 0.2, and 4.0 pp in table order. The

Table 4: **Factorial ablation over the high-level policy and the weight rule.** Success rate (%), 8 seeds for all ablation arms, mean±std. B0 is an ablation arm, not the published HIQL result. A0 is full NFTR. A3 omits the Bellman loss. A4 uses a frozen head and omits weight normalization. Per column, the best result is **orangebold**.

Arm	High-level policy	RWDR	Distance head	ant-tel-nav	ant-tel-stitch	pm-tel-nav	pm-tel-stitch	pm-med-stitch
B0	Gaussian	no	none	39.2 ± 1.9	31.3 ± 1.8	11.7 ± 0.4	26.7 ± 5.5	70.0 ± 6.0
G1	Gaussian	yes	co-trained	42.9 ± 0.7	35.5 ± 3.1	16.3 ± 3.0	31.8 ± 4.9	73.9 ± 10.1
G2	Gaussian	yes	frozen	40.1 ± 1.8	34.7 ± 2.7	13.2 ± 2.3	29.9 ± 5.7	76.1 ± 5.9
B1	NF	no	none	49.8 ± 0.3	40.2 ± 4.4	48.8 ± 5.4	39.3 ± 7.2	97.3 ± 2.7
A0	NF	yes	co-trained	**52.0 ± 0.6**	**44.0 ± 5.5**	**53.8 ± 5.9**	**41.7 ± 4.5**	**98.1 ± 1.2**
A2	NF	yes	frozen	50.0 ± 1.5	42.1 ± 1.8	49.9 ± 6.6	41.5 ± 9.6	94.1 ± 3.3
A3	NF, no Bellman term	yes	co-trained	49.6 ± 4.3	41.9 ± 1.5	49.2 ± 4.3	36.4 ± 2.9	95.6 ± 3.2
A4	NF, no weight normalization	yes	frozen	50.8 ± 2.1	40.1 ± 6.3	45.2 ± 4.2	39.1 ± 3.6	96.5 ± 2.2

0.2 pp difference on `pm-tel-stitch` is small relative to the reported variability. We therefore use A0 as the full method and treat A2 as a control. The comparison does not establish per-task significance or show that distance accuracy alone causes the gains.

Additional controls. Removing the Bellman term from the co-trained head (arm A3) lowers mean success relative to A0 on all five tasks, by 2.1 to 5.3 pp. Removing weight normalization from the frozen-head configuration (arm A4 versus A2) changes means in both directions. The latter comparison does not support a consistent gain from normalization.

Q4: How does the NF compare to a Gaussian across different configurations?

A4: Each matched NF arm (B1, A0, and A2) exceeds its Gaussian counterpart (B0, G1, and G2) in mean success across all five tasks. On `pm-tel-nav`, the three NF gains range from 36.7 to 37.5 pp. The gap remains under both co-trained and frozen-head RWDR, so the benefit of an expressive high-level policy is not limited to the value-only setting. These are matched configuration comparisons, not tests of significance (Table 4).

Table 5: **Selected RWDR coefficient results.** Success rate (%) on four tasks, 8 seeds, mean±std. The full sweep appears in Table 15. Per column, the highest mean is **orangebold** and the second-highest is *underlined*.

κ	ant-tel-nav	ant-tel-stitch	pm-med-stitch	cube-play
0.0	49.9 ± 0.3	41.1 ± 3.0	**98.1 ± 1.2**	22.2 ± 6.0
0.5	*50.5 ± 0.3*	40.5 ± 3.7	*98.0 ± 0.2*	24.7 ± 4.3
1.0	**52.0 ± 0.6**	**44.0 ± 5.5**	95.3 ± 2.9	31.3 ± 10.0
2.0	48.0 ± 3.1	*41.6 ± 1.6*	86.4 ± 8.7	**47.1 ± 9.9**
5.0	41.9 ± 3.2	33.1 ± 3.4	87.9 ± 5.5	*46.2 ± 4.4*

Q5: How sensitive is performance to the RWDR coefficient κ?

A5: The four tasks in Table 5 have different highest sample means. Both AntMaze teleport tasks favor $\kappa = 1$, `pm-med-stitch` favors 0, and `cube-play` favors 2. On `pm-med-stitch`, the mean falls from 98.1 at $\kappa = 0$ to 86.4 at $\kappa = 2$. The full 13-task sweep in Table 15 also favors 0 on `pm-tel-stitch`, 2 on `pm-tel-nav`, and 5 on `antmaze-teleport-explore`. Thus a stronger slack penalty does not give a monotonic gain, and $\kappa = 0$ remains useful when the NF alone captures the main benefit.

5 Conclusion

NFTR combines an NF high-level policy and triangle-slack reweighting to address HIQL's mode collapse and optimistic bias. It admits a three-term suboptimality bound (Theorem 3.8) and inherits AWR's population-level monotonic improvement (Proposition 3.9). On OGBench, the flow supplies most of the gain, while the full co-trained configuration adds task-dependent improvements over the value-only NF.

Limitations and Future Work. Our Gaussian analysis diagnoses mode averaging, while the RWDR guarantee assumes weighted maximum likelihood. Reweighting is task-dependent: its gains on a Gaussian backbone are smaller than those from switching to an NF, and some deterministic tasks favor $\kappa = 0$. On very long-horizon (`antmaze-giant-stitch`) and high-DoF contact-rich locomotion (`humanoidmaze`), planning and the physical realizability of NF-sampled modes become the bottlenecks. Temporal-abstraction methods like OTA perform better. The link between triangle-slack and stochastic invalidity is mechanistic rather than a bound. Combining NFTR with temporal abstraction is a natural next step.

AI USE STATEMENT

We used generative AI tools to assist with research brainstorming, manuscript drafting, and language polishing. All AI-assisted content was reviewed and revised by the authors, and all technical claims, experimental results, equations, and citations were independently verified. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work proposes an offline goal-conditioned reinforcement learning method and does not involve human subjects, personal data collection, or the release of new datasets. All experiments are conducted on publicly available benchmarks. Potential risks are limited to those generally associated with advances in machine learning, and we encourage responsible use consistent with standard safety and privacy practices.

REFERENCES

Hongjoon Ahn, Heewoong Choi, Jisu Han, and Taesup Moon. Option-aware temporally abstracted value for offline goal-conditioned reinforcement learning. *arXiv preprint arXiv:2505.12737*, 2025.

Dmitry Akimov, Vladislav Kurenkov, Alexander Nikulin, Denis Tarasov, and Sergey Kolesnikov. Let offline RL flow: Training conservative agents in the latent space of normalizing flows. In *3rd Offline RL Workshop: Offline RL as a "Launchpad"*, 2022.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

Seungho Baek, Taegeon Park, Jongchan Park, Seungjun Oh, and Yusung Kim. Graph-assisted stitching for offline hierarchical reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2391–2408. PMLR, 2025.

Marco Bagatella, Mert Albaba, Jonas Hübotter, Georg Martius, and Andreas Krause. Test-time offline reinforcement learning on goal-related experience. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025.

Janaka Brahmanage, Jiajing Ling, and Akshat Kumar. Flowpg: action-constrained policy gradient with normalizing flows. *Advances in Neural Information Processing Systems*, 36:20118–20132, 2023.

Chen-Hao Chao, Chien Feng, Wei-Fang Sun, Cheng-Kuang Lee, Simon See, and Chun-Yi Lee. Maximum entropy reinforcement learning via energy-based normalizing flow. *Advances in Neural Information Processing Systems*, 37:56136–56165, 2024.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017.

Zhiqiang Dong, Teng Pang, Rongjian Xu, and Guoqiang Wu. Efficient hierarchical implicit flow q-learning for offline goal-conditioned reinforcement learning. *arXiv preprint arXiv:2604.08960*, 2026.

Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Russ R Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. *Advances in Neural Information Processing Systems*, 35:35603–35620, 2022.

Benjamin Eysenbach, Vivek Myers, Ruslan Salakhutdinov, and Sergey Levine. Inference via interpolation: Contrastive representations provably enable planning and inference. *Advances in Neural Information Processing Systems*, 37:58901–58928, 2025.

Shaswat Garg, Matin Moezzi, and Brandon Da Silva. Data-efficient hierarchical goal-conditioned reinforcement learning via normalizing flows. *arXiv preprint arXiv:2602.11142*, 2026.

Raj Ghugare and Benjamin Eysenbach. Normalizing flows are capable models for rl. *arXiv preprint arXiv:2505.23527*, 2025.

Vittorio Giammarino, Ruiqi Ni, and Ahmed H Qureshi. Physics-informed value learner for offline goal-conditioned reinforcement learning. *arXiv preprint arXiv:2509.06782*, 2025.

Dan Haramati, Carl Qi, Tal Daniel, Amy Zhang, Aviv Tamar, and George Konidaris. Hierarchical entity-centric reinforcement learning with factored subgoal diffusion. In *International Conference on Learning Representations (ICLR)*, 2026.

Mineui Hong, Minjae Kang, and Songhwai Oh. Diffused task-agnostic milestone planner. *Advances in Neural Information Processing Systems*, 36:387–405, 2023.

Vineet Jain and Siamak Ravanbakhsh. Learning to reach goals via diffusion. In *International Conference on Machine Learning*, pp. 21170–21195. PMLR, 2024.

Hyungkyu Kang, Byeongchan Kim, and Min hwan Oh. Latent representation alignment for offline goal-conditioned reinforcement learning. In *Forty-third International Conference on Machine Learning*, 2026.

Kaiqiang Ke, Qian Lin, Zongkai Liu, Shenghong He, and Chao Yu. Conservative offline goal-conditioned implicit V-learning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29591–29607. PMLR, 2025.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2022.

Kyowoon Lee and Jaesik Choi. State-covering trajectory stitching for diffusion planners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Xing Lei, Wenyan Yang, Kaiqiang Ke, Shentao Yang, Xuetao Zhang, Joni Pajarinen, and Donglin Wang. Gchr: Goal-conditioned hindsight regularization for sample-efficient reinforcement learning. *arXiv preprint arXiv:2508.06108*, 2025.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Bo Liu, Yihao Feng, Qiang Liu, and Peter Stone. Metric residual network for sample efficient goal-conditioned reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8799–8806, 2023.

Yunhao Luo, Utkarsh A Mishra, Yilun Du, and Danfei Xu. Generative trajectory stitching through diffusion composition. *arXiv preprint arXiv:2503.05153*, 2025.

Yecheng Jason Ma, Jason Yan, Dinesh Jayaraman, and Osbert Bastani. How far i'll go: Offline goal-conditioned reinforcement learning via f-advantage regression. *arXiv preprint arXiv:2206.03023*, 2022.

Vivek Myers, Chongyi Zheng, Anca Dragan, Sergey Levine, and Benjamin Eysenbach. Learning temporal distances: Contrastive successor features can provide a metric structure for decision-making. *arXiv preprint arXiv:2406.17098*, 2024.

Vivek Myers, Catherine Ji, and Benjamin Eysenbach. Horizon generalization in reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2025a.

Vivek Myers, Bill Zheng, Benjamin Eysenbach, and Sergey Levine. Offline goal-conditioned reinforcement learning with quasimetric representations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Vivek Myers, Bill Chunyuan Zheng, Benjamin Eysenbach, and Sergey Levine. Offline goal-conditioned reinforcement learning with quasimetric representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025c.

Evgenii Opryshko, Junwei Quan, Claas Voelcker, Yilun Du, and Igor Gilitschenski. Test-time graph search for goal-conditioned reinforcement learning. *arXiv preprint arXiv:2510.07257*, 2025.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Seohong Park, Dibya Ghosh, Benjamin Eysenbach, and Sergey Levine. HIQL: Offline goal-conditioned RL with latent states as actions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Seohong Park, Tobias Kreiman, and Sergey Levine. Foundation policies with hilbert representations. *arXiv preprint arXiv:2402.15567*, 2024.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. Ogbench: Benchmarking offline goal-conditioned rl. In *International Conference on Learning Representations (ICLR)*, 2025a.

Seohong Park, Qiyang Li, and Sergey Levine. Flow q-learning. *arXiv preprint arXiv:2502.02538*, 2025b.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Moritz Reuss, Maximilian Li, Xiaogang Jia, and Rudolf Lioutikov. Goal-conditioned imitation learning using score-based diffusion policies. *arXiv preprint arXiv:2304.02532*, 2023.

Harshit Sikchi, Rohan Chitnis, Ahmed Touati, Alborz Geramifard, Amy Zhang, and Scott Niekum. Score models for offline goal-conditioned reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Avi Singh, Huihan Liu, Gaoyue Zhou, Albert Yu, Nicholas Rhinehart, and Sergey Levine. Parrot: Data-driven behavioral priors for reinforcement learning. *arXiv preprint arXiv:2011.10024*, 2020.

Mianchu Wang, Rui Yang, Xi Chen, Hao Sun, Meng Fang, and Giovanni Montana. GOPlan: Goal-conditioned offline reinforcement learning by planning with learned models. In *International Conference on Learning Representations (ICLR)*, 2025.

Tongzhou Wang, Antonio Torralba, Phillip Isola, and Amy Zhang. Optimal goal-reaching reinforcement learning via quasimetric learning. In *International Conference on Machine Learning*, pp. 36411–36430. PMLR, 2023.

Patrick Nadeem Ward, Ariella Smofsky, and Avishek Joey Bose. Improving exploration in soft-actor-critic with normalizing flows policies. *arXiv preprint arXiv:1906.02771*, 2019.

Haoran Xu, Li Jiang, Li Jianxiong, and Xianyuan Zhan. A policy-guided imitation approach for offline reinforcement learning. *Advances in neural information processing systems*, 35:4085–4098, 2022.

Haoran Xu, Shuozhe Li, Harshit Sikchi, Scott Niekum, and Amy Zhang. An optimal discriminator weighted imitation perspective for reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2025.

Youngsik Yoon, Gangbok Lee, Sungsoo Ahn, and Jungseul Ok. Breadth-first exploration on adaptive grid for reinforcement learning. In *Forty-first International Conference on Machine Learning*, 2024.

Bill Zheng, Vivek Myers, Benjamin Eysenbach, and Sergey Levine. Scaling goal-conditioned reinforcement learning with multistep quasimetric distances. In *The Fourteenth International Conference on Learning Representations*, 2026.

John L Zhou and Jonathan C Kao. Flattening hierarchies with policy bootstrapping. *arXiv preprint arXiv:2505.14975*, 2025.

Appendix

A Related Work

Offline Goal-Conditioned RL. Offline GCRL aims to learn goal-reaching policies from static datasets without environment interaction. Existing approaches can be categorized into several paradigms: goal-conditioned hindsight relabeling (Andrychowicz et al., 2017; Ke et al., 2025; Wang et al., 2025), hierarchical or subgoal-based learning (Park et al., 2023; Ahn et al., 2025; Giammarino et al., 2025; Zhou & Kao, 2025; Lei et al., 2025; Haramati et al., 2026), graph-based planning (Yoon et al., 2024; Eysenbach et al., 2025; Baek et al., 2025; Luo et al., 2025), metric learning (Wang et al., 2023; Park et al., 2024; Myers et al., 2024; 2025b;c;a), dual optimization (Ma et al., 2022; Sikchi et al., 2024; Xu et al., 2025), generative modeling (Hong et al., 2023; Reuss et al., 2023; Jain & Ravanbakhsh, 2024; Lee & Choi, 2025; Haramati et al., 2026), and test-time adaptation (Opryshko et al., 2025; Bagatella et al., 2025). HIQL (Park et al., 2023) extracts hierarchical policies from a single value function, achieving state-of-the-art performance on many tasks. However, HIQL relies on value-only subgoal selection, which can be overly optimistic in stochastic settings; our method augments subgoal learning with a geometric consistency signal to mitigate this issue.

Quasimetric Learning and Stochastic Environments. Quasimetric RL (Wang et al., 2023) establishes that the distance induced by an optimal goal-conditioned value function, $d^*(s, g) = -\log V^*(s, g)$, is a quasimetric under deterministic dynamics, an assumption the authors make explicit in their main text. The Metric Residual Network (MRN) (Liu et al., 2023) enforces this quasimetric structure architecturally. Contrastive Metric Distillation (CMD) (Myers et al., 2024) shows that contrastive temporal distances retain the triangle inequality even under stochastic dynamics. Building on this, Temporal Metric Distillation (TMD) (Myers et al., 2025b) is the first to systematically address quasimetric distance learning in offline GCRL with suboptimal, stochastic data, by combining Monte Carlo contrastive learning with a quasimetric architecture to recover optimal distances. More recently, Multistep Quasimetric Estimation (MQE) (Zheng et al., 2026) scales quasimetric learning by fitting distances with multistep Monte-Carlo returns, combining TD's local optimality with MC's global propagation. NFTR differs from this line in two aspects. First, we treat the quasimetric not as the policy-extraction backbone but as a *geometric scorer* that reweights the AWR objective inside HIQL's hierarchy, leaving value learning untouched. Second, we use the triangle inequality *as a regularizer on the high-level policy* via the slack penalty, which targets optimistic bias from stochastic transitions rather than the distance-estimation error itself.

Normalizing Flows in RL. Normalizing Flows enable exact density computation through invertible transformations (Dinh et al., 2014; 2017; Kingma & Dhariwal, 2018). In RL, NFs have been used for policy representation (Ward et al., 2019; Singh et al., 2020), maximum entropy RL (Chao et al., 2024), and action-constrained optimization (Brahmanage et al., 2023; Akimov et al., 2022). Flow Q-Learning (FQL) and its hierarchical variant (HFQL) (Park et al., 2025b) use flow-matching policies with distillation for stable value-guided offline RL, and NFs (Ghugare & Eysenbach, 2025) show NFs can serve as Q-functions or policies on broad benchmarks. Under weighted-MLE extraction, conditional NFs combine multi-modal density modeling, exact likelihood evaluation, and direct sampling. Weighted-regression alternatives do not require exact likelihoods and can also model the tilted subgoal target (Section 3.1 and Appendix F.3).

Hierarchical Offline RL. Beyond HIQL, recent hierarchical offline GCRL methods tackle long-horizon challenges via distinct strategies: SAW (Zhou & Kao, 2025) flattens the hierarchy via advantage-weighted subgoal sampling, OTA (Ahn et al., 2025) contracts effective horizons through temporal abstraction, and Eik-HiQRL (Giammarino et al., 2025) enforces quasimetric structure via Eikonal PDE constraints. While these reshape the hierarchy's *structure* through flattening, temporal abstraction, or PDE constraints, we leave HIQL's structure intact and instead address a property all share but none has examined, namely the behavior of AWR-weighted MLE on multi-modal subgoal data with stochastic advantages. Our contributions therefore compose with such structural improvements rather than competing with them.

Flow policies inside hierarchical offline GCRL. Two methods also place expressive flow policies inside HIQL. One uses Normalizing Flow policies at both hierarchy levels (Garg et al., 2026). The other uses a one-step mean-flow generator at both levels with a LeJEPA goal encoder (Dong et al.,

2026), and it appeared shortly before our deadline, so it is concurrent. Three differences separate them from NFTR. The first is placement, since we put the flow at the high level alone and Table 3 shows that a second flow costs 2.6 to 8.3 points. The second is motivation, since Theorem 3.1 derives the model class from the objective rather than adopting it for expressivity. The third is content, since neither method reweights the candidate set. Appendix F.1 reports the comparison.

B NOTATION

Table 6: Summary of additional notation used in the appendix (proofs, algorithm, and extended results). Symbols already defined in the main text (Sections 2 and 3) are not repeated.

Symbol	Description	Ref.
Multi-modal Structure and AWR Tilting (proof of Theorem 3.1)		
$M(s,g),\ \pi_m(s,g),$ $q_m(z\|s,g)$	State-dependent mode count; mode weights and conditional mode densities	Equation (6)
$\tilde{\pi}_m,\ \tilde{q}_m,\ \tilde{\mu}_m$	AWR-tilted mode weights, densities, and centers	Theorem 3.1
$p^\star_{s,g}(z),\ Z_{s,g},\ \beta_m(s,g)$	AWR target density; global and mode-wise normalizers	Theorem 3.1
$\mathcal{S}_{\text{wall}} \subset \mathbb{R}^d$	Unreachable region in subgoal space	Theorem 3.1
$\mu_\theta(s,g),\ \sigma$	Mean and isotropic std of the Gaussian high-level policy	Theorem 3.1
$A^H(s,w,g)$	Generic high-level advantage (instantiated as A_{env} in NFTR)	Equation (7)
Optimization and Loss Components (algorithm and Equation (17))		
α_L	Low-level AWR temperature	Algorithm 1
$\tau,\ L^2_\tau(u)$	IQL expectile parameter; asymmetric expectile loss	Equation (2)
τ_{ema}	Target-network EMA coefficient	Algorithm 1
$m = -\log\gamma$	Minimum one-step distance cost in Bellman consistency	Appendix C.1
$\Delta_{\max}$	Slack clipping threshold for numerical stability	Section 3.4
$\mathcal{L}_V,\ \mathcal{L}_{\pi^L}$	IQL value loss; low-level AWR loss	Equations (2) and (17)
$\mathcal{L}_{\text{NCE}},\ \mathcal{L}_{\text{Bell}},\ \mathcal{L}_d$	Contrastive, Bellman, and total distance losses	Appendix C.1
$\mathcal{L}^{\text{AWR}},\ \mathcal{L}_{\text{RWDR}}$	Generic AWR objective; RWDR weighted-MLE objective	Equations (7) and (12)
$g^+,\ g^-$	Positive / negative goal pairs for contrastive distance learning	Appendix C.1
Theoretical Quantities (Theorem 3.8 and Proposition 3.9)		
$\pi^{H,*},\ \hat{\pi}^H,\ \pi^{H,*}_{\mathcal{D}}$	Optimal, learned, and data-optimal high-level policies	Theorem 3.8 and Proposition 3.9
$J(\pi),\ J_{\text{RWDR}}(\pi)$	True expected return; RWDR objective value	Theorem 3.8
$A_{\max},\ D_{\max}$	Uniform bounds on $\|A_{\text{env}}\|$ and d_θ	Theorem 3.8
$R_\kappa(d_\theta)$	Geometric regularization term in the suboptimality bound	Theorem 3.8
$\Delta^*(s,w,g)$	Triangle-slack under the optimal distance d^*	Proposition 3.3
$\tilde{p}(z\|s,g)$	RWDR-weighted empirical target distribution	Proposition 3.9
$n,\ \|\Pi\|,\ d^\pi(s,w)$	Sample size; policy-class complexity; visitation distribution under π	Assumption 3.6 and Theorem 3.8
$C_\pi,\ \epsilon_V$	Concentrability coefficient; value estimation error bound	Assumptions 3.6 and 3.7

Algorithm 1 NFTR Training

Input: Offline dataset $\mathcal{D}$, hyperparameters $\alpha_H, \alpha_L, \kappa, \lambda_d, \lambda_{\text{Bell}}, \tau, \tau_{\text{ema}}$
Initialize: Value networks V_1, V_2, target networks $V_1^{\text{tgt}}, V_2^{\text{tgt}}$, goal representation ϕ, low-level actor π^L, high-level NFs π^H_{NF}, distance network d_θ
for each training iteration **do**
  Sample batch from $\mathcal{D}$: $\{(s_t, a_t, s_{t+1}, g_{\text{value}}, g_{\text{actor}}, w)\}_{i=1}^B$
  // Value Learning (IQL-style expectile regression)
  $q \leftarrow r + \gamma \cdot \min(V_1^{\text{tgt}}(s', g), V_2^{\text{tgt}}(s', g))$
  $\mathcal{L}_V \leftarrow \frac{1}{B}\sum_{i=1}^B \left[L_\tau^2(q_i - V_1(s_i, g_i)) + L_\tau^2(q_i - V_2(s_i, g_i))\right]$
  // Low-Level Actor (AWR with value advantage)
  $A_i^L \leftarrow V(s'_i, g_i^L) - V(s_i, g_i^L)$
  $\mathcal{L}_{\pi^L} \leftarrow -\frac{1}{B}\sum_{i=1}^B \exp(\alpha_L A_i^L) \cdot \log \pi^L(a_i | s_i, \phi([s_i; g_i^L]))$
  $\mathcal{L}_d \leftarrow 0$
  if $\lambda_d > 0$ **then**
    {full NFTR (A0); skipped for frozen-head A2}
    // Distance Learning (Contrastive + Bellman, see Appendix C.1)
    Sample positive goals g^+ from future states, negative goals g^- from other trajectories
    $\mathcal{L}_{\text{NCE}} \leftarrow -\frac{1}{B}\sum_{i=1}^B \log \frac{\exp(-d_\theta(s_i, g_i^+))}{\exp(-d_\theta(s_i, g_i^+)) + \exp(-d_\theta(s_i, g_i^-))}$
    $\mathcal{L}_{\text{Bell}} \leftarrow \frac{1}{B}\sum_{i=1}^B \left(\max(0, d_\theta(s_i, g_i) - d_\theta(s'_i, g_i) + \log\gamma)\right)^2$
    $\mathcal{L}_d \leftarrow \mathcal{L}_{\text{NCE}} + \lambda_{\text{Bell}}\mathcal{L}_{\text{Bell}}$
  end if
  // High-Level Actor (RWDR with NFs policy)
  $A_{\text{env},i} \leftarrow \texttt{stop_gradient}(V(w_i, g_i) - V(s_i, g_i))$
  $\Delta_i \leftarrow \max(0, \texttt{stop_gradient}(d_\theta(s_i, w_i) + d_\theta(w_i, g_i) - d_\theta(s_i, g_i)))$
  $\Delta_i \leftarrow \min(\Delta_i, \Delta_{\max})$ // Slack clipping for stability
  $\log \omega_i \leftarrow \alpha_H \cdot A_{\text{env},i} - \kappa \cdot \Delta_i$
  $\omega_i \leftarrow \min(\exp(\log \omega_i), 100)$ // Exponential clipping
  $\omega_i \leftarrow \omega_i / (\frac{1}{B}\sum_{j=1}^B \omega_j + \epsilon)$ // Weight normalization
  $z_i \leftarrow \text{stopgrad}\big(\phi([s_i; w_i])\big)$ {w_i is the waypoint state and ω_i is the scalar weight}
  $\epsilon_{\text{NF},i}, \log|\det J_i| \leftarrow \pi^H_{\text{NF}}.\text{forward}(s_i, g_i, z_i)$
  $\log \pi_i^H \leftarrow \log \mathcal{N}(\epsilon_{\text{NF},i}; 0, I) + \log|\det J_i|$
  $\mathcal{L}_{\pi^H} \leftarrow -\frac{1}{B}\sum_{i=1}^B \omega_i \cdot \log \pi_i^H$
  // Total Loss and Update
  $\mathcal{L} \leftarrow \mathcal{L}_V + \mathcal{L}_{\pi^L} + \mathcal{L}_{\pi^H} + \lambda_d \mathcal{L}_d$
  Update all parameters via gradient descent on $\mathcal{L}$
  $V^{\text{tgt}} \leftarrow (1 - \tau_{\text{ema}})V^{\text{tgt}} + \tau_{\text{ema}} V$ // Target network EMA update
end for
return π^H_{NF}, π^L

C Algorithm Details

This section spells out the training-time and test-time machinery summarized in Section 3.4. We first describe the quasimetric distance learning that underlies the triangle slack signal, then present the complete per-iteration training pseudocode, and finally describe the inference procedure used at evaluation.

C.1 Quasimetric Distance Learning

We parameterize the distance function d_θ with a Metric Residual Network (MRN) (Liu et al., 2023),

$$d_\theta(s, g) \;=\; \|\phi_{\text{sym}}(s) - \phi_{\text{sym}}(g)\|_2 \;+\; \text{ReLU}\Big(\max_i\big(\phi^{(i)}_{\text{asym}}(s) - \phi^{(i)}_{\text{asym}}(g)\big)\Big)\,, \tag{18}$$

where $\phi_{\text{sym}}, \phi_{\text{asym}} : \mathcal{S} \to \mathbb{R}^d$ are learned embeddings. The symmetric term provides a Euclidean component, while the asymmetric max-ReLU term captures directionality for one-way transitions.

By construction this architecture satisfies the triangle inequality $d_\theta(s, g) \leq d_\theta(s, w) + d_\theta(w, g)$ for all s, w, g, which is exactly the condition required by Assumption 3.4 and which makes the slack $\Delta(s, w, g)$ in Equation (10) non-negative.

We train d_θ via two complementary losses. The contrastive loss takes positive pairs (s, g^+) from the same trajectory with g^+ sampled geometrically from future states, and negative pairs (s, g^-) from different trajectories,

$$\mathcal{L}_{\text{NCE}} = -\mathbb{E}\left[\log \frac{\exp(-d_\theta(s, g^+))}{\exp(-d_\theta(s, g^+)) + \exp(-d_\theta(s, g^-))}\right]. \tag{19}$$

The Bellman consistency loss enforces that distances respect one-step transition dynamics,

$$\mathcal{L}_{\text{Bell}} = \mathbb{E}\left[\Big(\max\big(0,\ d_\theta(s, g) - d_\theta(s', g) - m\big)\Big)^2\right], \tag{20}$$

where $m = -\log \gamma$ is the minimum one-step cost. A one-step transition satisfies $d(s, g) \leq m + d(s', g)$, so only the violating direction is penalized. The total distance loss is $\mathcal{L}_d = \mathcal{L}_{\text{NCE}} + \lambda_{\text{Bell}} \mathcal{L}_{\text{Bell}}$, which is the term entering Equation (17) in the main text. Architectural non-negativity also holds for frozen-head A2. Co-trained A0 has higher mean success on all five tasks in Table 4. Full NFTR trains the MRN through both distance losses. These comparisons do not isolate distance-estimation accuracy.

C.2 Training Pseudocode

The complete NFTR training procedure is given in Algorithm 1. The algorithm jointly trains the value functions, the distance function d_θ, the low-level actor π^L, and the high-level Normalizing Flow policy π^H_{NF}. Per iteration, value, distance, and policy losses are evaluated on a single shared minibatch, and the high-level policy uses the RWDR weight (Equation (11)) computed from A_{env} and Δ with stop-gradient on V_θ and d_θ. The stop-gradient on the high-level target z_i prevents the high-level likelihood loss from directly updating ϕ. It does not freeze ϕ across iterations: the value update, and any other loss with an open gradient path to the shared encoder, can move ϕ and hence the latent subgoal targets seen by the NF. The fixed-snapshot arguments in Appendices D.1 and D.4 do not bound this temporal drift.

C.3 Inference Procedure

At test time, given the current state s and goal g, we generate a subgoal and execute it via the following steps:

1. Sample $\epsilon \sim \mathcal{N}(0, I)$ from the NF base distribution.
2. Compute the subgoal representation $z = f_\theta^{-1}(\epsilon;\ [s; g])$ via the inverse NF.
3. Apply the test-time length normalization $z \leftarrow z \cdot \|\phi([s; g])\|_2 / (\|z\|_2 + \varepsilon)$.
4. Execute the low-level policy $a = \pi^L(s, z)$ for k steps, or until the subgoal is reached.
5. Repeat from step 1 with the updated state.

This rescaling is an inference heuristic: the NF training targets $\phi([s; w])$ are not length-normalized. It therefore changes the scale of sampled subgoals relative to the training distribution, and the theoretical analysis does not cover its effect. The reported evaluation uses this inference procedure; a matched training-time normalization or an ablation of the heuristic would be needed to isolate its contribution.

D Proofs

We provide complete proofs for all theoretical results stated in the main text. The proofs establish the geometric properties of triangle-slack, convergence guarantees for RWDR, robustness to stochastic optimism, and monotonic improvement.

D.1 Proof of Theorem 3.1

Setup. Fix (s, g) and write $A^H(z) := A^H(s, w(z), g)$ for brevity, where $w(z)$ is the deterministic preimage of $z = \phi([s; w])$ along the sampled trajectory. By the bounded-and-integrable assumption on $e^{\alpha_H A^H}$,

$$Z_{s,g} \;:=\; \mathbb{E}_{z\sim p_{\mathcal{D}}(\cdot|s,g)}\big[e^{\alpha_H A^H(z)}\big] \;\in\; (0,\infty),$$

so the AWR target density $p^\star_{s,g}(z) = Z_{s,g}^{-1}\, e^{\alpha_H A^H(z)}\, p_{\mathcal{D}}(z \mid s, g)$ is a well-defined probability density on $\mathbb{R}^d$ that retains the support of $p_{\mathcal{D}}$.

Step 1: closed-form minimizer. Expand $\log \mathcal{N}(z; \mu, \sigma^2 I_d) = -\frac{1}{2\sigma^2}\|z-\mu\|_2^2 + C(\sigma)$, where $C(\sigma) = -\frac{d}{2}\log(2\pi\sigma^2)$ is independent of μ. Substituting into Equation (7) gives, up to terms that do not depend on μ_θ,

$$\mathcal{L}^{\mathrm{AWR}}(\theta; s, g) \;=\; \frac{1}{2\sigma^2}\,\mathbb{E}_{z\sim p_{\mathcal{D}}(\cdot|s,g)}\Big[e^{\alpha_H A^H(z)}\,\|z-\mu_\theta(s,g)\|_2^2\Big] \;+\; \text{const.}$$

The change of measure $\mathbb{E}_{p_{\mathcal{D}}}[e^{\alpha_H A^H} f(z)] = Z_{s,g}\,\mathbb{E}_{p^\star_{s,g}}[f(z)]$, valid for any integrable f, then yields

$$\mathcal{L}^{\mathrm{AWR}}_{\mu}(\theta; s, g) \;=\; \frac{Z_{s,g}}{2\sigma^2}\,\mathbb{E}_{z\sim p^\star_{s,g}}\big[\|z-\mu_\theta(s,g)\|_2^2\big],$$

which is strictly convex in $\mu_\theta(s, g)$ since $\sigma > 0$ and $Z_{s,g} > 0$. The first-order condition

$$\nabla_{\mu_\theta(s,g)}\mathcal{L}^{\mathrm{AWR}}_{\mu} \;=\; -\,\frac{Z_{s,g}}{\sigma^2}\,\mathbb{E}_{z\sim p^\star_{s,g}}\big[z-\mu_\theta(s,g)\big] \;=\; 0$$

admits the unique solution $\mu^\star_\theta(s, g) = \mathbb{E}_{z\sim p^\star_{s,g}}[z]$.

Step 2: mixture decomposition of $p^\star_{s,g}$. Substituting Equation (6),

$$\begin{aligned} p^\star_{s,g}(z) \;&=\; Z_{s,g}^{-1}\, e^{\alpha_H A^H(z)} \sum_{m=1}^{M(s,g)} \pi_m(s,g)\, q_m(z \mid s, g) \\ &=\; \sum_{m=1}^{M(s,g)} \underbrace{\frac{\pi_m(s,g)\,\beta_m(s,g)}{Z_{s,g}}}_{=:\,\tilde\pi_m(s,g)} \cdot \underbrace{\frac{e^{\alpha_H A^H(z)}\, q_m(z \mid s, g)}{\beta_m(s,g)}}_{=:\,\tilde q_m(z|s,g)}, \end{aligned}$$

where $\beta_m(s, g) := \mathbb{E}_{z\sim q_m(\cdot|s,g)}[e^{\alpha_H A^H(z)}] \in (0, \infty)$. Each $\tilde q_m$ is a normalized density on $\mathbb{R}^d$, namely the AWR-tilted version of q_m. The weights $\tilde\pi_m$ are non-negative and sum to one, since

$$\sum_{m=1}^{M(s,g)} \pi_m(s,g)\,\beta_m(s,g) \;=\; \mathbb{E}_{z\sim p_{\mathcal{D}}(\cdot|s,g)}\big[e^{\alpha_H A^H(z)}\big] \;=\; Z_{s,g}.$$

Therefore $p^\star_{s,g} = \sum_{m=1}^{M(s,g)} \tilde\pi_m\, \tilde q_m$ is a valid mixture.

Step 3: AWR-tilted modal centers. Define $\tilde\mu_m(s, g) := \mathbb{E}_{z\sim\tilde q_m(\cdot|s,g)}[z]$. Direct computation gives

$$\tilde\mu_m(s,g) \;=\; \frac{\mathbb{E}_{z\sim q_m(\cdot|s,g)}\big[e^{\alpha_H A^H(z)}\, z\big]}{\mathbb{E}_{z\sim q_m(\cdot|s,g)}\big[e^{\alpha_H A^H(z)}\big]}.$$

Since $q_m(\cdot \mid s, g)$ is supported on a path-connected region disjoint from $\mathcal{S}_{\text{wall}}$ and AWR tilting preserves support, $\tilde q_m$ is supported on the same region, so $\tilde\mu_m(s, g) \in \mathrm{conv}\big(\mathrm{supp}(q_m(\cdot \mid s, g))\big)$.

Step 4: convex combination. Combining Steps 1 to 3, by linearity of expectation,

$$\mu^\star_\theta(s,g) \;=\; \mathbb{E}_{z\sim p^\star_{s,g}}[z] \;=\; \sum_{m=1}^{M(s,g)} \tilde\pi_m(s,g)\,\mathbb{E}_{z\sim\tilde q_m(\cdot|s,g)}[z] \;=\; \sum_{m=1}^{M(s,g)} \tilde\pi_m(s,g)\,\tilde\mu_m(s,g),$$

which establishes Equation (8). Since $\tilde\pi_m \geq 0$ and $\sum_m \tilde\pi_m = 1$, the right-hand side is a convex combination, hence $\mu^\star_\theta(s, g) \in \mathrm{conv}\{\tilde\mu_m(s, g)\}_{m=1}^{M(s,g)}$.

Step 5: when the convex combination falls inside $\mathcal{S}_{\text{wall}}$. Assume $\mathcal{S}_{\text{wall}} \subset \mathbb{R}^d$ strictly separates the supports $\{\text{supp}(q_m)\}_m$ into two non-empty groups, indexed by G_1 and G_2. Let

$$p_j \;=\; \sum_{m \in G_j} \tilde{\pi}_m(s,g), \qquad \bar{\mu}_j \;=\; \sum_{m \in G_j} \frac{\tilde{\pi}_m(s,g)}{p_j}\, \tilde{\mu}_m(s,g), \qquad j \in \{1,2\}.$$

By Step 4, $\mu^\star_\theta(s,g) = p_1\bar{\mu}_1 + p_2\bar{\mu}_2$, with $\bar{\mu}_1$ and $\bar{\mu}_2$ on opposite sides of $\mathcal{S}_{\text{wall}}$ by construction. The line segment $[\bar{\mu}_1, \bar{\mu}_2]$ then crosses $\mathcal{S}_{\text{wall}}$ at an interior point $p^\star$. Whenever the AWR-tilted weights produce a (p_1, p_2) that matches the convex coefficient of $p^\star$ on this segment, $\mu^\star_\theta(s,g)$ lies inside $\mathcal{S}_{\text{wall}}$. The symmetric two-mode configuration $M = 2$, $\tilde{\pi}_1 = \tilde{\pi}_2 = \frac{1}{2}$ realizes this generically when $\bar{\mu}_1$ and $\bar{\mu}_2$ are reflectionally symmetric about $\mathcal{S}_{\text{wall}}$, which is the canonical two-corridor case in our experiments.

Step 6: irreducible wall mass on any practical scale. Suppose the configuration in Step 5 places $\mu^\star_\theta(s,g)$ in $\mathcal{S}_{\text{wall}}$. Then there exists an open ball $B_r(\mu^\star_\theta) \subset \mathcal{S}_{\text{wall}}$ for some $r > 0$, and

$$\mathbb{P}_{z \sim \mathcal{N}(\mu^\star_\theta, \sigma^2 I_d)}\big[z \in B_r(\mu^\star_\theta)\big] \;=\; \mathbb{P}_{u \sim \mathcal{N}(0, I_d)}\big[\|u\|_2 \le r/\sigma\big].$$

This probability tends to 1 as $\sigma \to 0^+$ and is strictly positive for every finite $\sigma > 0$, since the standard Gaussian assigns positive mass to every ball of positive radius. Within any fixed compact range $\sigma \in [\sigma_{\min}, \sigma_{\max}]$ that arises in practice,

$$\inf_{\sigma \in [\sigma_{\min}, \sigma_{\max}]} \mathbb{P}_{z \sim \mathcal{N}(\mu^\star_\theta, \sigma^2 I_d)}\big[z \in \mathcal{S}_{\text{wall}}\big] \;\ge\; \mathbb{P}_{u \sim \mathcal{N}(0, I_d)}\big[\|u\|_2 \le r/\sigma_{\max}\big] \;>\; 0,$$

so no choice of σ in this range drives the wall-mass to zero. The degenerate limit $\sigma \to \infty$ is excluded because the AWR weights would then fail the bounded-and-integrable assumption used in Step 1.

D.2 Proof of Proposition 3.3 (Slack Characterization)

Proof. We prove each part of the proposition separately.

Non-negativity. By Definition 1, any quasimetric d satisfies the triangle inequality:

$$d(s,g) \le d(s,w) + d(w,g), \quad \forall s, w, g \in \mathcal{S}. \tag{21}$$

Rearranging this inequality:

$$d(s,w) + d(w,g) - d(s,g) \ge 0, \quad \forall s, w, g \in \mathcal{S}. \tag{22}$$

Since the triangle-slack is defined as $\Delta(s,w,g) = \max(0, d(s,w) + d(w,g) - d(s,g))$ and we have shown the argument of the max is non-negative, we conclude:

$$\Delta(s,w,g) = d(s,w) + d(w,g) - d(s,g) \ge 0. \tag{23}$$

Geodesic characterization. We prove both directions of the if-and-only-if.

($\Rightarrow$) Suppose $\Delta(s,w,g) = 0$. Then:

$$\max(0, d(s,w) + d(w,g) - d(s,g)) = 0. \tag{24}$$

Since the argument is non-negative by Part 1, this implies:

$$d(s,w) + d(w,g) - d(s,g) = 0 \implies d(s,g) = d(s,w) + d(w,g). \tag{25}$$

By Definition 3, w is geodesic for (s,g) under d.

($\Leftarrow$) Suppose w is geodesic for (s,g), i.e., $d(s,g) = d(s,w) + d(w,g)$ by Definition 3. Then:

$$d(s,w) + d(w,g) - d(s,g) = 0. \tag{26}$$

Therefore:

$$\Delta(s,w,g) = \max(0,0) = 0. \tag{27}$$

Detour characterization. This follows directly from Part 2 by contraposition.

($\Rightarrow$) If $\Delta(s,w,g) > 0$, then by Part 2, w is not geodesic, meaning $d(s,g) < d(s,w) + d(w,g)$. The path $s \to w \to g$ has strictly greater total distance than the direct path $s \to g$, which defines a detour.

($\Leftarrow$) If reaching g via w requires a detour, then by definition $d(s,w) + d(w,g) > d(s,g)$. Thus:

$$\Delta(s,w,g) = d(s,w) + d(w,g) - d(s,g) > 0. \tag{28}$$

Optimal distance correspondence (deterministic MDP). Define $d^*(s,g) := -\log V^*(s,g)$, where $V^*(s,g) \in (0,1]$ is the optimal discounted reaching probability. The goal-reaching Bellman optimality equation is

$$V^*(s,g) = \mathbf{1}[s=g] + \gamma \max_a \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s',g)]. \tag{29}$$

In a deterministic MDP, fix a state w on an optimal trajectory from s to g. Let $\pi^*_{s\to g}$ denote an optimal policy for goal g starting from s; by determinism this policy reaches g along a single trajectory $s = s_0, s_1, \ldots, s_{k_1} = w, \ldots, s_{k_1+k_2} = g$ that passes through w at step k_1 and through g at step $k_1 + k_2$. Hence

$$V^*(s,g) \;=\; \gamma^{k_1+k_2} \;=\; \gamma^{k_1} \cdot \gamma^{k_2} \;=\; V^*(s,w) \cdot V^*(w,g). \tag{30}$$

Taking $-\log$ on both sides, $d^*(s,g) = d^*(s,w) + d^*(w,g)$, which by Definition 3 means w is geodesic for (s,g). Conversely, if w is not on any optimal trajectory from s to g, then any policy that reaches g via w takes strictly more than $k_1 + k_2$ steps in expectation (or fails to reach g at all), so $V^*(s,w) \cdot V^*(w,g) < V^*(s,g)$ and hence $d^*(s,g) < d^*(s,w) + d^*(w,g)$, i.e., $\Delta^*(s,w,g) > 0$. This establishes that geodesic subgoals under d^* are exactly those on optimal paths in the deterministic case.

Stochastic MDPs (triangle inequality survives, geodesic correspondence relaxes). In a general MDP with stochastic transitions, multiplicative equality fails because reaching w no longer uniquely determines the conditional distribution over subsequent trajectories. We show that the triangle inequality survives by an explicit policy-concatenation argument. Let $\pi_{s\to w}$ be any policy that, starting from s, reaches w with discounted probability $V^{\pi_{s\to w}}(s,w)$, and let $\pi_{w\to g}$ be any policy that, starting from w, reaches g with discounted probability $V^{\pi_{w\to g}}(w,g)$. Define the concatenated policy $\pi_{s\to w} \circ \pi_{w\to g}$: follow $\pi_{s\to w}$ until reaching w, then switch to $\pi_{w\to g}$. By the strong Markov property and the multiplicativity of discount factors,

$$V^{\pi_{s\to w}\circ\pi_{w\to g}}(s,g) \;=\; V^{\pi_{s\to w}}(s,w) \cdot V^{\pi_{w\to g}}(w,g). \tag{31}$$

Since $\pi_{s\to w} \circ \pi_{w\to g}$ is a feasible (in general suboptimal) policy for the $s \to g$ problem,

$$V^*(s,g) \;\geq\; V^{\pi_{s\to w}\circ\pi_{w\to g}}(s,g) \;=\; V^{\pi_{s\to w}}(s,w) \cdot V^{\pi_{w\to g}}(w,g). \tag{32}$$

Taking the supremum over $\pi_{s\to w}$ and $\pi_{w\to g}$ separately on the right-hand side (the two suprema decouple because the factors depend on disjoint policy components),

$$V^*(s,g) \;\geq\; V^*(s,w) \cdot V^*(w,g), \tag{33}$$

which after applying $-\log$ yields the triangle inequality

$$d^*(s,g) \;\leq\; d^*(s,w) + d^*(w,g). \tag{34}$$

Thus $\Delta^*(s,w,g) \geq 0$ remains well-defined and Parts (1)-(3) of the proposition apply verbatim with $d = d^*$.

The deterministic-case characterization "$\Delta^* = 0$ iff w is on an optimal path", however, no longer holds in both directions. Equality in Equation (34) can occur even for off-path w: if a stochastic transition from s produces unusually favorable outcomes that boost $V^*(s,w)$ above the average reachability of w from s, then $V^*(s,w) \cdot V^*(w,g)$ may match $V^*(s,g)$ without w being on any optimal trajectory. We refer to such transitions as *lucky* in the main text. Conversely, geodesic subgoals on optimal paths still satisfy $\Delta^* = 0$ because Equation (31) attains $V^*(s,g)$ when both factors are realized by the optimal policy. Triangle-slack thus continues to lower-bound composability failure: $\Delta^*(s,w,g) > 0$ implies w is off-path, but $\Delta^*(s,w,g) = 0$ no longer guarantees on-path.

The MRN is trained separately from V_θ on trajectory pairs and one-step transitions. Its architecture guarantees non-negative slack. Its learned geometry can assign larger detours to unreliable waypoints, an association tested in Appendix F.5. This evidence does not imply that every lucky transition has positive slack. □

D.3 Proof of Corollary 3.5 (Geometric Regularization Effect)

Proof. By the definition of RWDR weights:

$$w_{\text{RWDR}}(s,w,g) = \exp(\alpha_H A_{\text{env}}(s,w,g) - \kappa\Delta(s,w,g)) \tag{35}$$
$$= \exp(\alpha_H A_{\text{env}}(s,w,g)) \cdot \exp(-\kappa\Delta(s,w,g)) \tag{36}$$
$$= w_{\text{HIQL}}(s,w,g) \cdot \exp(-\kappa\Delta(s,w,g)). \tag{37}$$

Since the MRN architecture guarantees $d_\theta(s,g) \le d_\theta(s,w) + d_\theta(w,g)$, we have $\Delta(s,w,g) \ge 0$ for all (s,w,g). Therefore:

$$\exp(-\kappa\Delta(s,w,g)) \le 1, \tag{38}$$

with strict inequality when $\Delta(s,w,g) > 0$ and $\kappa > 0$. This implies:

$$w_{\text{RWDR}}(s,w,g) \le w_{\text{HIQL}}(s,w,g), \tag{39}$$

with strict inequality when both $\kappa > 0$ and $\Delta(s,w,g) > 0$. □

D.4 Proof of Theorem 3.8 (RWDR Convergence)

Proof. We decompose the suboptimality gap into three interpretable terms and bound each separately.

Performance Decomposition Let $\pi^{H,*}$ denote the optimal high-level policy and $\hat{\pi}^H$ denote the policy learned via RWDR. We decompose the suboptimality gap:

$$J(\pi^{H,*}) - J(\hat{\pi}^H) = \underbrace{[J(\pi^{H,*}) - J_{\text{RWDR}}(\pi^{H,*})]}_{\text{(I) Objective gap}} \tag{40}$$
$$+ \underbrace{[J_{\text{RWDR}}(\pi^{H,*}) - J_{\text{RWDR}}(\hat{\pi}^H)]}_{\text{(II) Optimization gap}} \tag{41}$$
$$+ \underbrace{[J_{\text{RWDR}}(\hat{\pi}^H) - J(\hat{\pi}^H)]}_{\text{(III) Transfer gap}}, \tag{42}$$

where $J(\pi)$ is the true expected return and $J_{\text{RWDR}}(\pi)$ is the RWDR objective value.

Bounding Term (I) - Value Estimation Error The RWDR objective uses estimated advantages $\hat{A}_{\text{env}}(s,w,g) = \hat{V}_\theta(w,g) - \hat{V}_\theta(s,g)$. Under Assumption 3.7:

$$\mathbb{E}_{(s,g)\sim\mathcal{D}}[(V_\theta(s,g) - V^*(s,g))^2] \le \epsilon_V. \tag{43}$$

By the triangle inequality for the L^2 norm, we have:

$$|\hat{A}_{\text{env}}(s,w,g) - A^*_{\text{env}}(s,w,g)| = |(\hat{V}(w,g) - V^*(w,g)) - (\hat{V}(s,g) - V^*(s,g))| \tag{44}$$
$$\le |\hat{V}(w,g) - V^*(w,g)| + |\hat{V}(s,g) - V^*(s,g)|. \tag{45}$$

Taking expectations on both sides of $|\hat{A}_{\text{env}} - A^*_{\text{env}}| \le |\hat{V}(w,g) - V^*(w,g)| + |\hat{V}(s,g) - V^*(s,g)|$ and applying Cauchy-Schwarz $\mathbb{E}[|X|] \le \sqrt{\mathbb{E}[X^2]}$ to each term,

$$\mathbb{E}_{\mathcal{D}}[|\hat{A}_{\text{env}} - A^*_{\text{env}}|] \;\le\; 2\sqrt{\epsilon_V}. \tag{46}$$

To compare the RWDR targets induced by estimated and true advantages, fix a snapshot of the co-trained distance head d_θ during policy extraction. Write the target induced by an advantage estimate $\hat{A}$ as

$$p_{\hat{A}}(w \mid s,g) \;\propto\; \beta(w \mid s,g) \cdot \exp\big(\alpha_H \hat{A}(s,w,g) - \kappa\Delta(s,w,g)\big), \tag{47}$$

and write p_A for the same object built from the true advantage A. Both targets use the same slack function Δ from this fixed, trained snapshot; stop-gradient prevents the policy loss from updating the head, while $\mathcal{L}_d$ trains it separately. Hence

$$\frac{p_{\hat{A}}(w)}{p_A(w)} \;=\; \exp\big(\alpha_H(\hat{A} - A)\big) \cdot \frac{Z_A}{Z_{\hat{A}}}. \tag{48}$$

The geometric factor cancels pointwise in the unnormalized ratio, but the normalizers still depend on κ and d_θ through the reweighted base measure. Let $h_{\hat{A}} = \exp(\alpha_H \hat{A} - \kappa\Delta)$ and $h_A = \exp(\alpha_H A - \kappa\Delta)$. Since both advantages have magnitude at most $A_{\max}$ and $0 \le \Delta \le 2D_{\max}$, their normalizers are at least $\exp(-\alpha_H A_{\max} - 2\kappa D_{\max})$, and $|h_{\hat{A}} - h_A| \le \alpha_H \exp(\alpha_H A_{\max})|\hat{A} - A|$. Thus, conditionally on (s, g),

$$\|p_{\hat{A}} - p_A\|_1 \;\le\; 2\alpha_H \exp(2\alpha_H A_{\max} + 2\kappa D_{\max})\, \mathbb{E}_{\beta(w|s,g)}[|\hat{A} - A|]. \tag{49}$$

Under Assumption 3.7, averaging the right-hand side over the high-level sampling distribution and applying the Cauchy–Schwarz bound above controls the value-estimation contribution with a constant that may depend on κ and $D_{\max}$, rather than independently of the distance head:

$$\text{Term (I)} \le \mathcal{O}_{\alpha_H, \kappa, A_{\max}, D_{\max}}(\sqrt{\epsilon_V}). \tag{50}$$

What the geometric factor does change is the location of the target. This is a displacement rather than an amplification, and it is bounded by construction, since $0 \le \Delta \le 2D_{\max}$ gives $\mathrm{KL}(p_{\text{RWDR}} \,\|\, p_{\text{AWR}}) \le 2\kappa D_{\max}$. This is why the geometric quantity enters Theorem 3.8 as the additive term $R_\kappa(d_\theta)$ rather than as a multiplier on ϵ_V. No assumption on the quality of d_θ enters at any point, since $\Delta \ge 0$ holds for every parameter setting of the MRN head.

Bounding Term (II) - Optimization Error The RWDR objective is a weighted maximum likelihood:

$$\mathcal{L}_{\text{RWDR}}(\pi^H) = -\mathbb{E}_{(s,w,g)\sim\mathcal{D}}[w(s, w, g) \cdot \log \pi^H(z|s, g)]. \tag{51}$$

Under Assumption 3.6, the behavior policy β covers the support of the optimal policy $\pi^{H,*}$ with density ratio bounded by C_π. By standard concentration inequalities for importance-weighted empirical risk minimization (Rashidinejad et al., 2021):

$$|J_{\text{RWDR}}(\pi^{H,*}) - J_{\text{RWDR}}(\hat{\pi}^H)| \le \mathcal{O}\left(\sqrt{\frac{C_\pi \log |\Pi|}{n}}\right), \tag{52}$$

where $|\Pi|$ is the complexity of the policy class (which can be replaced by Rademacher complexity for continuous function classes via standard generalization bounds).

Regularization Term Analysis (Bridging $\exp(-\kappa\Delta)$ to $R_\kappa(d_\theta)$).

The RWDR weight contains the geometric penalty factor $\exp(-\kappa\Delta(s, w, g))$. We now derive how this multiplicative factor on the per-sample weight translates into the additive regularization term $R_\kappa(d_\theta)$ in the suboptimality bound, completing the missing intermediate steps.

Step 1 (Bias of the RWDR-weighted expected advantage). The RWDR objective extracts a policy whose AWR target distribution downweights samples by $\exp(-\kappa\Delta)$. Define the RWDR-weighted expected advantage of policy π as

$$\widetilde{A}_{\text{RWDR}}(\pi) \;:=\; \mathbb{E}_{(s,w,g)\sim\pi}\big[A_{\text{env}}(s, w, g) \cdot \exp(-\kappa\Delta(s, w, g))\big], \tag{53}$$

and let $\widetilde{A}(\pi) := \mathbb{E}_{(s,w,g)\sim\pi}[A_{\text{env}}(s, w, g)]$ denote the unweighted version. Then by linearity of expectation,

$$\widetilde{A}(\pi) - \widetilde{A}_{\text{RWDR}}(\pi) \;=\; \mathbb{E}_{(s,w,g)\sim\pi}\big[A_{\text{env}}(s, w, g) \cdot (1 - \exp(-\kappa\Delta(s, w, g)))\big], \tag{54}$$

so the bias introduced by the geometric reweighting is controlled by the per-sample factor $1 - \exp(-\kappa\Delta) \in [0, 1]$.

Step 2 (Linearization-free upper bound). Since $\Delta(s, w, g) \ge 0$ by Assumption 3.4, we have $\exp(-\kappa\Delta) \in (0, 1]$, hence $1 - \exp(-\kappa\Delta) \in [0, 1]$. By the relation $J_{\text{RWDR}}(\pi) - J(\pi) = \widetilde{A}_{\text{RWDR}}(\pi) - \widetilde{A}(\pi)$ established in Step 1 and the triangle inequality for expectations,

$$\big|J_{\text{RWDR}}(\pi) - J(\pi)\big| \;\le\; \mathbb{E}_{(s,w,g)\sim\pi}[|A_{\text{env}}(s, w, g)| \cdot (1 - \exp(-\kappa\Delta(s, w, g)))]\,. \tag{55}$$

Applying the elementary inequality $1 - \exp(-x) \le x$ for all $x \ge 0$, which holds globally rather than only as a first-order Taylor approximation,

$$\big|J_{\text{RWDR}}(\pi) - J(\pi)\big| \;\le\; \mathbb{E}_{(s,w,g)\sim\pi}[|A_{\text{env}}(s, w, g)| \cdot \kappa\,\Delta(s, w, g)]\,. \tag{56}$$

This bound is tight to first order in $\kappa\Delta$ and holds globally for any $\Delta \ge 0$.

Step 3 (Reduction to an additive regularization term). Specializing the bound to the optimal high-level policy $\pi^{H,*}$ and absorbing the bounded advantage $|A_{\text{env}}| \le A_{\max}$ into a constant:

$$\left|J_{\text{RWDR}}(\pi^{H,*}) - J(\pi^{H,*})\right| \;\le\; \kappa A_{\max} \cdot \mathbb{E}_{(s,w,g)\sim\pi^{H,*}}[\Delta(s,w,g)] \;=:\; R_\kappa(d_\theta). \tag{57}$$

This identifies $R_\kappa(d_\theta)$ as the additive regularization term in the final bound.

Step 4 (Bounded magnitude regardless of training). Since the MRN architecture ensures $d_\theta(s,g) \le D_{\max}$ for bounded state spaces, $\Delta(s,w,g) \le d_\theta(s,w) + d_\theta(w,g) \le 2D_{\max}$. Therefore:

$$R_\kappa(d_\theta) \;\le\; 2\kappa A_{\max} D_{\max}, \tag{58}$$

Step 5 (Vanishing regularization for trained d_θ). When d_θ approximates optimal distances on $\text{supp}(\pi^{H,\star})$, geodesic subgoals satisfy $\Delta \approx 0$ by Proposition 3.3(2), so $R_\kappa(d_\theta) \to 0$.

Combining the three terms. Summing the value-error, sampling-error, and regularization contributions gives the bound stated in Theorem 3.8.

$$J(\pi^{H,\star}) - J(\hat{\pi}^H) \le \mathcal{O}(\sqrt{\epsilon_V}) + \mathcal{O}\left(\sqrt{\frac{C_\pi \log|\Pi|}{n}}\right) + R_\kappa(d_\theta).$$

□

D.5 Proof of Proposition 3.9 (Monotonic Improvement)

Proof. We show that the RWDR objective inherits the monotonic improvement property of weighted maximum likelihood estimation.

Identify the Target Distribution The RWDR objective (Equation (12)) has the form:

$$\mathcal{L}_{\pi^H} = -\mathbb{E}_{(s,w,g)\sim\mathcal{D}}[w(s,w,g) \cdot \log \pi^H(\phi([s;w])|s,g)], \tag{59}$$

where the RWDR weight $w(s,w,g) = \exp(\alpha_H A_{\text{env}}(s,w,g) - \kappa\Delta(s,w,g)) \ge 0$ is non-negative for all (s,w,g).

Define the weighted empirical distribution over subgoal representations:

$$\tilde{p}(z|s,g) \propto \sum_{w\in\mathcal{D}_{s,g}} w(s,w,g) \cdot \delta(\phi([s;w]) - z), \tag{60}$$

where $\mathcal{D}_{s,g} = \{w : (s,w,g) \in \mathcal{D}\}$ is the set of subgoals observed for the state-goal pair (s,g) in the dataset. The normalization ensures $\int \tilde{p}(z|s,g)dz = 1$.

The optimal in-distribution high-level policy is:

$$\pi^{H,*}_{\mathcal{D}}(z|s,g) = \tilde{p}(z|s,g). \tag{61}$$

Weighted MLE Minimizes KL Divergence The negative log-likelihood objective under weighting w is:

$$\mathcal{L}_{\pi^H} = -\mathbb{E}_{(s,w,g)\sim\mathcal{D}}[w(s,w,g) \cdot \log \pi^H(\phi([s;w])|s,g)]. \tag{62}$$

This is equivalent to minimizing the KL divergence from the target distribution to the policy:

$$\text{KL}(\tilde{p}\|\pi^H) = \mathbb{E}_{z\sim\tilde{p}(z|s,g)}[\log\tilde{p}(z|s,g) - \log\pi^H(z|s,g)] \tag{63}$$

$$= \underbrace{\mathbb{E}_{z\sim\tilde{p}}[\log\tilde{p}(z|s,g)]}_{\text{constant w.r.t. } \pi^H} - \mathbb{E}_{z\sim\tilde{p}}[\log\pi^H(z|s,g)]. \tag{64}$$

Since the first term is constant with respect to π^H, minimizing $\mathcal{L}_{\pi^H}$ is equivalent to minimizing $\text{KL}(\tilde{p}\|\pi^H)$.

Monotonic Decrease of KL Divergence (Population Level). The negative log-likelihood objective $\mathcal{L}_{\pi^H}$ is, up to an additive constant, equal to $\text{KL}(\tilde{p}\|\pi^H)$. At the population level, by definition of the exact minimizer $\pi^H_{\text{new}} = \arg\min_{\pi^H\in\Pi}\mathcal{L}_{\pi^H}$ over the policy class Π, for any $\pi^H_{\text{old}} \in \Pi$,

$$\text{KL}(\tilde{p}\|\pi^H_{\text{new}}) \;\le\; \text{KL}(\tilde{p}\|\pi^H_{\text{old}}). \tag{65}$$

This is the population-level statement of Proposition 3.9. We emphasize that this argument is about the exact minimizer over Π rather than about any iterative procedure, and we discuss the practical SGD optimization separately below. Specifically, for any policy π^H_{old}, a gradient step yields π^H_{new} satisfying:

$$\text{KL}(\tilde{p}\|\pi^H_{\text{new}}) \leq \text{KL}(\tilde{p}\|\pi^H_{\text{old}}). \tag{66}$$

Since $\tilde{p} = \pi^{H,*}_{\mathcal{D}}$, we have:

$$\text{KL}(\pi^{H,*}_{\mathcal{D}}\|\pi^H_{\text{new}}) \leq \text{KL}(\pi^{H,*}_{\mathcal{D}}\|\pi^H_{\text{old}}). \tag{67}$$

Relation to Expected Return The target distribution $\pi^{H,*}_{\mathcal{D}}$ places higher probability mass on subgoals with:

- High value advantage $A_{\text{env}}(s, w, g)$: subgoals that make progress toward the goal
- Low triangle-slack $\Delta(s, w, g)$: subgoals that are geometrically on-path

By decreasing $\text{KL}(\pi^{H,*}_{\mathcal{D}}\|\pi^H)$, the policy π^H increasingly assigns probability to these high-quality subgoals. Within the support of the dataset (the in-distribution region), this corresponds to improving expected return.

The key insight is that RWDR modifies the weights in AWR but preserves the weighted maximum likelihood structure. Since the weights remain non-negative and the objective is still weighted log-likelihood, standard AWR convergence analysis applies at the population level, establishing monotonic improvement of $\text{KL}(\pi^{H,*}_{\mathcal{D}}\|\pi^H)$.

Practical caveat (SGD on non-convex parameterization). The result above assumes exact optimization of the weighted MLE objective at the population level. In practice, π^H is a conditional RealNVP optimized by SGD, for which the exact-minimizer inequality does not imply per-step improvement. This optimization gap also arises in other AWR-based methods and is not specific to RWDR. The evaluation success curves in Appendix F.10 show stabilization in representative tasks, but they do not measure the weighted-MLE loss or establish convergence to a stationary point. □

E Hyperparameter Settings

We provide complete hyperparameter settings for reproducing our experiments.

Table 7: Common hyperparameters shared across all tasks.

Hyperparameter	**Locomotion**	**Manipulation**
Training steps	1,000,000	1,000,000
Batch size	1024	1024
Optimizer	Adam	Adam
Learning rate	3×10^{-4}	3×10^{-4}
Discount factor γ	0.99	0.99
Target network EMA τ_{ema}	0.005	0.005
IQL expectile τ	0.7	0.7
Low-level temperature α_L	3.0	3.0
High-level temperature α_H	3.0	3.0
Subgoal horizon k	25	25
Goal representation dimension	10	10
Hidden layer dimension	256	256
Number of hidden layers	2	2

Table 8: NFTR-specific hyperparameters. The κ entries are reference settings. Task-specific settings used for reported NFTR results are identified in Appendix F.8.

Hyperparameter	**Stochastic/Stitch**	**Standard**
RWDR penalty coefficient κ	1.0	0.5
Slack clipping threshold $\Delta_{\max}$	10.0	10.0
Distance loss weight λ_d	1.0	1.0
Bellman consistency weight λ_{Bell}	1.0	1.0
NFs number of coupling layers K	4	4
NFs hidden dimension	256	256
NFs context embedding dimension	128	128
Distance network latent dimension	64	64
MRN number of components	8	8
Weight normalization	True	True
Exponential weight clipping	100.0	100.0

Table 9: Normalizing Flow architecture details.

Component	**Specification**
Base distribution	$\mathcal{N}(0, I)$ with dimension 10
Number of coupling layers	4
Coupling layer type	Affine (RealNVP-style)
Scale/translation network	2-layer MLP, 256 hidden units, ReLU
Conditioning mechanism	Concatenation of $[s; g]$ to each layer input
Permutation between layers	Alternating fixed permutation
Output activation	None (linear)

F Additional Experimental Results

F.1 Comparison with Hierarchical Flow Variants of HIQL

For the Normalizing Flow variant, we ran the authors' public implementation on the four state-based teleport datasets under our protocol. The mean-flow variant uses our reimplementation because no public implementation was available for this comparison; its numbers should be read with that caveat.

Table 10: **Two-level flow variants of HIQL against NFTR.** Success rate (%) on the four state-based teleport datasets, mean±std. Our runs use 8 seeds. NF-HIQL is the published two-level Normalizing Flow method and not the ablation arm B1 of Table 4. HIFQL is our reimplementation of the two-level mean-flow method. HIQL values are the published ones. Per column, the best result is **orangebold**.

Method	`pm-tel-nav`	`pm-tel-stitch`	`ant-tel-nav`	`ant-tel-stitch`
HIQL	18.0 ± 4.0	34.0 ± 4.0	42.0 ± 3.0	36.0 ± 2.0
NF-HIQL	33.6 ± 5.7	35.8 ± 3.1	43.4 ± 1.6	41.5 ± 1.3
HIFQL	30.2 ± 4.6	36.4 ± 4.6	42.3 ± 4.5	31.6 ± 5.8
NFTR (ours)	**53.8 ± 5.9**	**41.7 ± 4.5**	**52.0 ± 0.6**	**44.0 ± 5.5**

Analysis. Both methods improve on HIQL where the Gaussian collapses hardest, moving 18.0 to 33.6 and to 30.2 on `pm-tel-nav`. Elsewhere the gain is small. NF-HIQL stays within 1.4 to 5.5 points of HIQL on the other three datasets, and HIFQL falls below HIQL on `ant-tel-stitch`. Both land in the same band, and NFTR is ahead on all four. A flow inserted into HIQL recovers part of the gap. This cross-method comparison cannot assign the remaining difference to one component. Table 4 provides the closer within-framework comparison.

F.2 Comparison with Value-Representation Methods

CGCIVL (Ke et al., 2025) and LAVL (Kang et al., 2026) improve the value or the representation that produces the advantage. RWDR acts later, at policy extraction, once an advantage is already available. We therefore also test whether the two compose. NF-LAVL below pairs the LAVL value representation with our NF high-level policy. Neither published method reports the stochastic teleport datasets, so the numbers are our own runs under the protocol of Table 1.

Table 11: **Value-representation baselines and one composed arm.** Success rate (%), mean±std, on the four state-based teleport datasets. Our runs use 8 seeds. NF-LAVL carries no reweighting. HIQL values are the published ones. Per column, the best result is **orangebold**.

Method	`pm-tel-nav`	`pm-tel-stitch`	`ant-tel-nav`	`ant-tel-stitch`
HIQL	18.0 ± 4.0	34.0 ± 4.0	42.0 ± 3.0	36.0 ± 2.0
CGCIVL	40.1 ± 6.2	38.6 ± 3.5	37.9 ± 1.8	35.4 ± 1.7
LAVL	21.3 ± 2.1	30.3 ± 2.8	47.0 ± 0.5	37.3 ± 1.8
NF-LAVL	23.4 ± 1.0	31.8 ± 7.9	46.1 ± 3.3	41.9 ± 2.1
NFTR (ours)	**53.8 ± 5.9**	**41.7 ± 4.5**	**52.0 ± 0.6**	**44.0 ± 5.5**

Analysis. Each baseline is strongest on one maze family and gives the gain back on the other. CGCIVL leads the two PointMaze columns among the value-representation baselines and sits 4.1 and 0.6 points below HIQL on AntMaze. LAVL leads the two AntMaze columns and sits 3.7 points below HIQL on `pm-tel-stitch`. NFTR is ahead in all four, by 13.7 and 3.1 points over CGCIVL and by 5.0 and 6.7 points over LAVL on the family where each is strongest. We do not claim the 3.1 on its own because the intervals overlap.

The composed arm cuts both ways. NF-LAVL improves on LAVL by $+2.1$, $+1.5$ and $+4.6$ points, loses 0.9 points on `ant-tel-nav`, and stays 2.1 to 30.4 points below NFTR. A stronger value representation does not make the flow redundant, since LAVL alone leads no column, and an expressive policy on a stronger value backbone does not reproduce NFTR either. On `pm-tel-nav`, B1 exceeds the ablation baseline B0 by 37.1 points, while NF-LAVL exceeds LAVL by 2.1 points. AWR reweights the subgoals proposed by each policy, so value learning and policy expressivity act at different stages.

Two structural reasons complete the answer. First, a value scores a pair and states how good w is, whereas the slack scores a triple and states whether $s \to w \to g$ is internally consistent, so no improvement to a pairwise value produces this quantity on its own. Second, the value cannot supply the residual. One could form $\Delta_V = -\log[V(s,w)V(w,g)/V(s,g)]$, but V_θ is fit by expectile regression on an action-free surrogate under no triangle constraint, so Δ_V takes both signs and $\exp(-\kappa\Delta_V)$ would reward a negative residual, which is exactly the shortcut the penalty should suppress. The MRN head is used because $\Delta \geq 0$ holds architecturally, which is what makes the penalty one-sided and conservative.

The scope of this comparison is the teleport datasets that NFTR targets. On the deterministic navigate and stitch suites both baselines report results we do not match, and we claim no general ranking.

F.3 Why Normalizing Flows: Comparison with Diffusion and Flow Matching

To address the question of *why* we choose Normalizing Flows over other expressive multi-modal generative families (diffusion models, flow matching), we compare NFTR against two natural alternatives that share the same hierarchy and triangle-slack reweighting:

- **FMTR** (Flow-Matching subgoal policy with Triangle-slack Reweighting): replaces the conditional NF head with a conditional flow-matching subgoal generator, sampled with a 20-step Euler integrator at inference.
- **DMTR** (Diffusion subgoal policy with Triangle-slack Reweighting): replaces the NF head with a conditional diffusion subgoal generator, sampled with 20-step DDIM.

Everything else (hierarchy, value learning, RWDR weighting, hyperparameters) is held fixed.

Table 12: **Expressive high-level heads under an identical hierarchy and weight.** Success rate (%), mean±std. Our runs use 8 seeds. HIQL values are from the original paper. FMTR replaces the NF head with a conditional flow-matching generator and DMTR with a conditional diffusion generator sampled by DDIM, while hierarchy, value learning, RWDR weight and hyperparameters are held fixed. Per row, the best result is **orangebold** and the second-best is *underlined*.

Task	HIQL	NFTR	FMTR	DMTR
pointmaze-teleport-navigate	18.0 ± 4.0	**53.8 ± 5.9**	38.5 ± 1.0	37.9 ± 5.7
pointmaze-teleport-stitch	34.0 ± 4.0	41.7 ± 4.5	**45.6 ± 0.9**	40.1 ± 6.3
pointmaze-medium-stitch	74.0 ± 6.0	**98.1 ± 1.2**	95.2 ± 0.8	96.2 ± 2.6
antmaze-teleport-navigate	42.0 ± 3.0	**52.0 ± 0.6**	44.1 ± 3.0	42.6 ± 4.6
antmaze-teleport-stitch	36.0 ± 2.0	**44.0 ± 5.5**	40.4 ± 0.9	36.2 ± 2.8
cube-single-play	15.0 ± 3.0	**47.1 ± 9.9**	35.6 ± 8.6	17.4 ± 4.5
scene-noisy	25.0 ± 4.0	32.6 ± 5.8	**42.9 ± 1.0**	31.4 ± 1.7

Analysis. Table 12 places NFTR ahead on five of the seven completed datasets, including margins of 15.3 points on `pointmaze-teleport-navigate` and 7.9 points on `antmaze-teleport-navigate`. FMTR leads on `pointmaze-teleport-stitch` by 3.9 points and on `scene-noisy` by 10.3 points. All three expressive heads improve on HIQL on every completed dataset.

FMTR and DMTR are trained by weighted regression rather than weighted maximum likelihood, so their training does not require exact density evaluation and (R2) does not apply to these arms. A one-step generator can also satisfy (R3), although the implementations compared here use iterative sampling. These alternatives show that expressive high-level subgoal modeling admits another extraction route; they do not alter the Gaussian mode-averaging diagnosis of Theorem 3.1, which also holds under weighted least-squares regression by Remark 3.2. RWDR acts on the sample weight and can be combined with either extraction route.

What remains for the flow are reasons rather than necessity. The first is the empirical margin above. The second is that the exact density is used directly in Figure 7, and the NF head adds no integrator and no step-count hyperparameter, whereas FMTR uses a 20-step Euler integrator and DMTR a 20-step DDIM schedule. The third is that Theorem 3.8 and Proposition 3.9 are formulated for weighted maximum likelihood. The population targets coincide, so the remaining margin can only reflect approximation, optimization or finite-sample effects, and we record it as an open question.

F.4 How the Distance Enters the High-Level Weight

Appendix F.2 compared value and distance signals. We now test how distance enters the high-level weight. All four arms use an NF high-level policy. The value-only arm has no distance head, while the other arms use distance terms in the weight.

Analysis. Two comparisons carry the argument. *Distance only* has a lower four-task mean than *Value only*, at 40.6 versus 44.5. *Value + distance progress* against NFTR is the matched comparison, since both arms use a value and a distance and differ only in how the distance enters. A progress term is maximized by whatever reduces $d_\theta(\cdot, g)$ fastest, and a teleporter does exactly that, so the term rewards the shortcut and gains nothing over the value alone, at 44.0 against 44.5. Before batch normalization, positive slack can only reduce the raw weight. NFTR reaches a four-task mean of 47.9. The essential point is that the sign constraint, rather than the distance itself, is what makes the geometric signal useful.

F.5 Direct Evidence That High-Slack Subgoals Are Unreliable

On NFTR models and four teleport datasets, we take every transition $s_t \to w = s_{t+25}$, reset the simulator to s_t and replay the identical action sequence. The dispersion of the replayed endpoints then measures non-repeatability on its own terms. No teleporter location and no learned label enters this measurement.

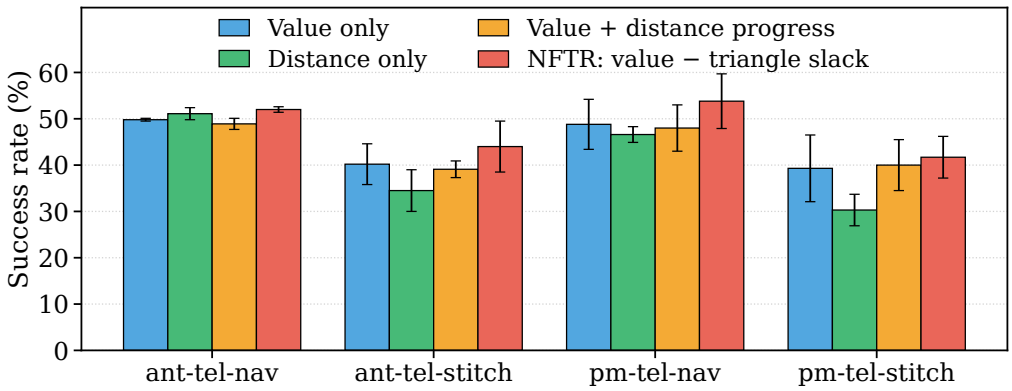


Figure 4: **Four weight rules on the same NF backbone.** Success rate (%) on the four teleport datasets, 8 seeds, error bars are one standard deviation. *Value only* is the plain AWR weight $\exp(\alpha_H A_{\text{env}})$ and coincides with arm B1 of Table 4. *Distance only* replaces the advantage by the directional progress $d_\theta(s, g) - d_\theta(w, g)$. *Value + distance progress* adds that progress term to the advantage. *NFTR* subtracts the triangle slack from the advantage. Averaged over the four datasets the arms reach 44.5, 40.6, 44.0 and 47.9, which the figure does not display.

Analysis. Teleport transitions are genuinely non-repeatable, since identical actions scatter their endpoints by roughly 15 units while ordinary transitions stay near zero. The slack ranks a non-repeatable transition above a repeatable one on all four datasets. The association also concentrates where it matters. The highest-slack decile is the most teleport-dense in every dataset, at 1.77 to 2.70 times the base rate, and restricted to apparently nearby subgoals AUROC reaches 0.781 and 0.806 on the two navigate datasets. The unconditional value is lower because most transitions are on-path regardless. RWDR converts this signal into a lower weight through $\exp(-\kappa \min(\Delta, \Delta_{\max}))$, and where the term is active a teleport subgoal receives roughly 31 to 34 percent less weight than a normal one at matched advantage. The essential point is that the association is measured on the environment rather than inferred from the value function, so it does not inherit the bias it is meant to correct.

F.6 Puzzle Manipulation Results

We evaluate NFTR on the OGBench Puzzle suite, which involves discrete grasp-and-press primitives that produce inherently multi-modal subgoal distributions.

Table 13: **OGBench Puzzle Results.** Success rate (%) on Puzzle 3×3, 4×4, 4×5, and 4×6, in both `play` and `noisy` settings. NFTR uses 8 seeds, mean±std. HIQL values are from the original paper. Per row, the best result is **orangebold** and the second-best is *underlined*.

Task	HIQL	NFTR
`puzzle-3x3-play`	12 ± 2	$\mathbf{14.9 \pm 2.2}$
`puzzle-4x4-play`	7 ± 2	$\mathbf{38.1 \pm 6.9}$
`puzzle-4x5-play`	4 ± 1	$\mathbf{6.7 \pm 1.0}$
`puzzle-4x6-play`	3 ± 1	$\mathbf{3.5 \pm 0.2}$
`puzzle-3x3-noisy`	$\mathbf{51 \pm 11}$	18.0 ± 2.3
`puzzle-4x4-noisy`	16 ± 4	$\mathbf{53.0 \pm 6.7}$
`puzzle-4x5-noisy`	5 ± 1	$\mathbf{8.8 \pm 0.8}$
`puzzle-4x6-noisy`	2 ± 1	$\mathbf{5.2 \pm 2.4}$

Analysis. Table 13 shows that NFTR achieves substantial gains over HIQL on the medium-grid configurations (`4x4-play`: +31.1 pp, `4x4-noisy`: +37.0 pp) where the multi-modal structure of valid grasp-and-press primitives matches the Normalizing Flow's modeling capacity, and where observation noise allows triangle-slack to filter spuriously-rewarding subgoals. As grid size grows to 4×5 and 4×6, the bottleneck shifts from per-step subgoal quality to long-horizon *combinatorial planning* (the number of feasible block-press sequences explodes), and the gains compress for both methods. The regression on `3x3-noisy` is substantial: NFTR scores 18.0 against HIQL's 51.0, a 33.0 pp decrease. These results alone do not establish the cause of that decrease. Overall, the

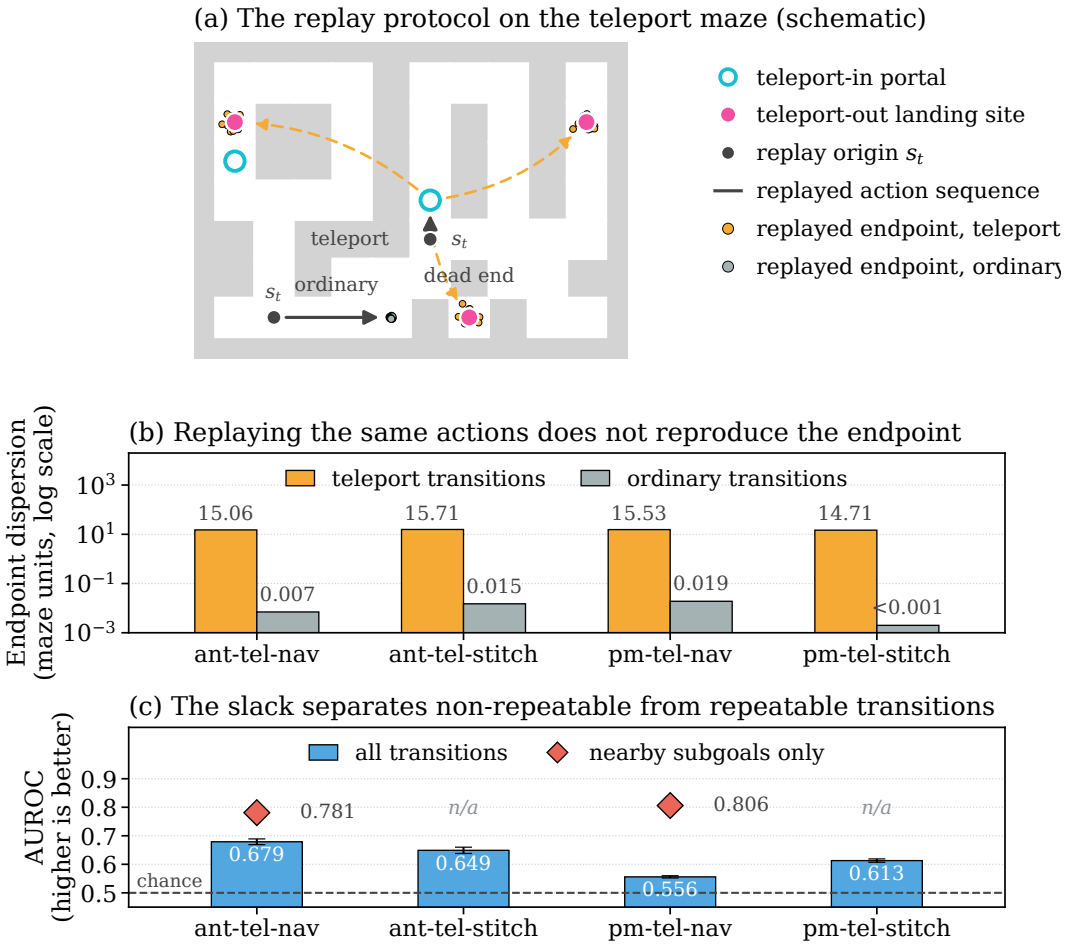


Figure 5: **Replay evidence that high-slack subgoals are non-repeatable. (a)** Schematic of the protocol on the `teleport` layout, drawn cell by cell from Figure 7 and using its marker convention. An ordinary transition returns the same endpoint every time. A transition that passes a teleport-in portal is dispatched to one of three teleport-out sites at random, and one of those sites is a dead end, so a single action sequence produces endpoints in three different rooms. **(b)** Measured endpoint dispersion on a logarithmic axis, in the coordinate units of the environment. Teleport transitions scatter by roughly 15 units and ordinary ones stay below 0.02, and the value for `pm-tel-stitch` falls below the axis floor and is annotated instead. **(c)** AUROC with which the slack ranks a non-repeatable transition above a repeatable one, over 8 seeds with ± 1 standard deviation, and the dashed line marks chance. Blue bars use all transitions. Red diamonds restrict the pool to apparently nearby subgoals, where a lucky shortcut is the actual failure mode, and that pool is defined only on the two navigate datasets, so the other two entries are marked as not available. Both panels use the co-trained head.

results suggest that NFTR helps most when (i) multi-modality is present and (ii) noise/stochasticity is non-trivial, while remaining orthogonal to long-horizon planning improvements.

F.7 HumanoidMaze Results: A Negative Result

We honestly report HumanoidMaze results, where NFTR underperforms HIQL.

Analysis. Table 14 reports a consistent regression of NFTR relative to HIQL across all four HumanoidMaze settings (e.g., -39.5 pp on `medium-navigate`, -41.3 pp on `large-navigate`). This regression has two coupled causes:

Table 14: **OGBench HumanoidMaze Results.** Success rate (%) on HumanoidMaze `medium`/`large` navigate and stitch settings (21-DoF humanoid, contact-rich dynamics). NFTR uses 8 seeds, mean±std. HIQL values are from the original paper. Per row, the best result is **orangebold** and the second-best is *underlined*.

Task	HIQL	NFTR
humanoidmaze-medium-navigate	**89 ± 2**	49.5 ± 3.8
humanoidmaze-large-navigate	**49 ± 4**	7.7 ± 1.5
humanoidmaze-medium-stitch	**88 ± 2**	62.9 ± 7.5
humanoidmaze-large-stitch	**28 ± 3**	7.8 ± 2.3

1. **Physical realizability of multi-modal subgoals.** The 21-DoF humanoid's contact-rich dynamics make many of the multiple subgoal modes generated by the Normalizing Flow not all physically realizable: a "valid corridor" entrance in maze coordinates may be unreachable for a humanoid given balance and footstep constraints. Consequently, multi-modal subgoal proposals introduce noise into AWR training, which a unimodal Gaussian (HIQL) avoids by collapsing to the single safest mode.
2. **Limited environmental stochasticity.** HumanoidMaze does not contain teleporters or strong random transitions, so the optimistic-bias issue that triangle-slack is designed to address is largely absent. RWDR therefore provides limited gain while still introducing additional optimization complexity.

This is consistent with our scope statement (Section 4). NFTR targets mode collapse and optimistic bias, and gains can arise when either is prominent. In high-DoF, contact-rich, deterministic regimes, neither appears to dominate. The additional NF and RWDR machinery does not overcome the long-horizon control difficulty here.

F.8 Comprehensive RWDR Coefficient κ Ablation

We report the full κ sweep across the 13 environments in Table 15. For these tasks, the NFTR scores reported elsewhere correspond to $\kappa = 0$ on `pm-tel-stitch`, `pm-med-stitch`, `antmaze-medium-stitch`, and `antmaze-large-stitch`. They correspond to $\kappa = 1$ on `pointmaze-large-stitch`, `ant-tel-nav`, and `ant-tel-stitch`, and to $\kappa = 2$ on `pm-tel-nav` and the four Cube and Scene tasks. The `antmaze-teleport-explore` result uses $\kappa = 5$. The values in Table 8 are reference settings, not a universal coefficient.

Table 15: **Full κ Ablation.** Sensitivity to the RWDR penalty coefficient $\kappa \in \{0, 0.5, 1.0, 2.0, 5.0\}$ across the 13 environments included in this sweep. Success rate (%), 8 seeds, mean±std. Per row (i.e., per environment), the best κ is **orangebold** and the second-best is *underlined*.

Task	κ=0.0	κ=0.5	κ=1.0	κ=2.0	κ=5.0
pointmaze-teleport-navigate	48.2 ± 3.6	48.7 ± 4.5	48.6 ± 10.5	**53.8 ± 5.9**	47.3 ± 2.5
pointmaze-teleport-stitch	**41.7 ± 4.5**	37.1 ± 9.1	40.8 ± 8.8	40.6 ± 7.9	40.0 ± 8.9
pointmaze-medium-stitch	**98.1 ± 1.2**	98.0 ± 0.2	95.3 ± 2.9	86.4 ± 8.7	87.9 ± 5.5
pointmaze-large-stitch	30.3 ± 6.3	29.6 ± 5.6	30.0 ± 5.7	**31.8 ± 5.9**	28.8 ± 4.9
antmaze-teleport-navigate	49.9 ± 0.3	50.5 ± 0.3	**52.0 ± 0.6**	48.0 ± 3.1	41.9 ± 3.2
antmaze-teleport-stitch	41.1 ± 3.0	40.5 ± 3.7	**44.0 ± 5.5**	41.6 ± 1.6	33.1 ± 3.4
antmaze-teleport-explore	14.2 ± 2.7	17.6 ± 4.5	19.8 ± 3.1	19.0 ± 1.4	**26.1 ± 1.4**
antmaze-medium-stitch	**81.0 ± 3.7**	80.8 ± 4.3	76.0 ± 5.5	70.2 ± 8.5	31.0 ± 6.2
antmaze-large-stitch	**16.9 ± 2.1**	13.6 ± 2.5	14.4 ± 2.4	15.9 ± 0.8	11.7 ± 2.1
cube-single-play	22.2 ± 6.0	24.7 ± 4.3	31.3 ± 10.0	**47.1 ± 9.9**	46.2 ± 4.4
cube-single-noisy	55.2 ± 3.3	62.8 ± 5.1	64.0 ± 2.7	**67.4 ± 0.9**	59.4 ± 3.9
scene-play	51.7 ± 3.7	**52.2 ± 1.1**	47.7 ± 8.7	45.6 ± 3.8	24.8 ± 5.0
scene-noisy	27.1 ± 2.1	32.8 ± 3.1	**34.8 ± 2.4**	32.6 ± 5.8	22.2 ± 4.7

Three regimes emerge from Table 15.

(1) Stitching tasks. The highest means on `pointmaze-medium-stitch` and `antmaze-medium-stitch` occur at $\kappa = 0$ in Table 15. On the former task, this configuration scores 98.1% versus HIQL's reported 74%. A positive slack penalty provides no higher mean on these two tasks.

(2) Teleport tasks. The highest mean occurs at $\kappa = 0$ for `pm-tel-stitch`, at 1 for the two AntMaze teleport tasks, at 2 for `pm-tel-nav`, and at 5 for `antmaze-teleport-explore`. The preferred coefficient varies by task.

(3) Manipulation tasks. The highest means occur at $\kappa = 2$ for both Cube tasks, at 0.5 for `scene-play`, and at 1 for `scene-noisy`. The preferred coefficient varies across these tasks.

The sweep shows that the preferred κ varies across tasks, with both gains and regressions as the penalty increases. It therefore describes the scope and sensitivity of RWDR rather than establishing a universal coefficient or a monotonic relation with environmental stochasticity.

F.9 Comparison with NF-RLBC Variants

For completeness, we compare NFTR against goal-conditioned and hierarchical goal-conditioned variants of NF-RLBC (Ghugare & Eysenbach, 2025), denoted **GC-NF-RLBC** and **H-GC-NF-RLBC**. These baselines also use Normalizing Flows without our triangle-slack reweighting. This comparison measures overall performance against NF-based alternatives. It does not isolate RWDR because the methods also differ in hierarchy and training objective.

Table 16: **Comparison with NF-RLBC variants.** Success rate (%), mean±std. Our runs use 8 seeds. HIQL values are from the original paper. GC-NF-RLBC: goal-conditioned NF-RLBC. H-GC-NF-RLBC: hierarchical goal-conditioned NF-RLBC. Per row, the best result is **orangebold** and the second-best is *underlined*.

Task	HIQL	NFTR	GC-NF-RLBC	H-GC-NF-RLBC
`pointmaze-teleport-navigate`	18 ± 4	**53.8 ± 5.9**	13.2 ± 6.4	*40.1 ± 6.5*
`pointmaze-teleport-stitch`	*34 ± 4*	**41.7 ± 4.5**	19.5 ± 2.8	4.9 ± 4.8
`pointmaze-medium-stitch`	*74 ± 6*	**98.1 ± 1.2**	22.3 ± 6.5	0.3 ± 0.4
`pointmaze-large-stitch`	*13 ± 6*	**30.0 ± 5.7**	2.0 ± 2.1	0.8 ± 1.5
`pointmaze-giant-stitch`	0 ± 0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
`antmaze-teleport-navigate`	42 ± 3	**52.0 ± 0.6**	33.2 ± 3.2	*49.5 ± 2.4*
`antmaze-teleport-stitch`	36 ± 2	**44.0 ± 5.5**	14.5 ± 1.3	*36.8 ± 2.0*
`antmaze-teleport-explore`	**34 ± 15**	*26.1 ± 1.4*	0.7 ± 0.3	0.1 ± 0.1
`antmaze-medium-stitch`	**94 ± 1**	*81.0 ± 3.7*	29.0 ± 3.2	61.0 ± 2.2
`antmaze-large-stitch`	**67 ± 5**	*16.9 ± 2.1*	0.2 ± 0.1	4.2 ± 2.3
`antmaze-giant-stitch`	**2 ± 2**	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
`cube-single-play`	15 ± 3	**47.1 ± 9.9**	*18.7 ± 0.3*	9.1 ± 2.5
`cube-single-noisy`	*41 ± 6*	**67.4 ± 0.9**	14.1 ± 0.8	18.3 ± 1.6
`scene-play`	*38 ± 3*	**45.6 ± 3.8**	10.2 ± 1.2	20.4 ± 2.6
`scene-noisy`	*25 ± 4*	**32.6 ± 5.8**	4.1 ± 0.2	11.3 ± 0.7

Analysis. Table 16 shows NFTR outperforming both GC-NF-RLBC and H-GC-NF-RLBC on 13 of 15 datasets and tying on the remaining two, where all three methods score zero. The flat goal-conditioned NF policy (GC-NF-RLBC) struggles broadly. The hierarchical variant (H-GC-NF-RLBC) recovers some performance, but NFTR has higher mean success on all 13 non-tied tasks, often by a substantial margin. The gap on `pointmaze-medium-stitch`, at 98.1 against 0.3, is not evidence for the reweighting. Table 15 reports 98.1 on this dataset at $\kappa = 0$, so the reweighting cannot be responsible for it. The two methods also differ in hierarchy, value learning, and policy training. This cross-method gap cannot be assigned to one component. On `antmaze-medium/large/giant-stitch`, HIQL itself outperforms NFTR, consistent with the limits of this comparison on long-horizon tasks.

F.10 Training Curves and Practical Stability under SGD

We record evaluation success every 0.1×10^6 gradient steps on the 15 evaluated tasks. Figure 6 shows four representative regimes. Mean success levels off after roughly 0.3 to 0.5×10^6 steps in most of these runs, without a collapse in the plotted interval. For panel (b), seed-level standard deviation contracts from ± 11.9 at 0.1×10^6 steps to ± 2.4 at 1×10^6 steps; panel (c) varies between ± 1.4 and ± 5.7 over training. These curves describe evaluation behavior. They do not establish per-step improvement or convergence of the weighted-MLE objective.

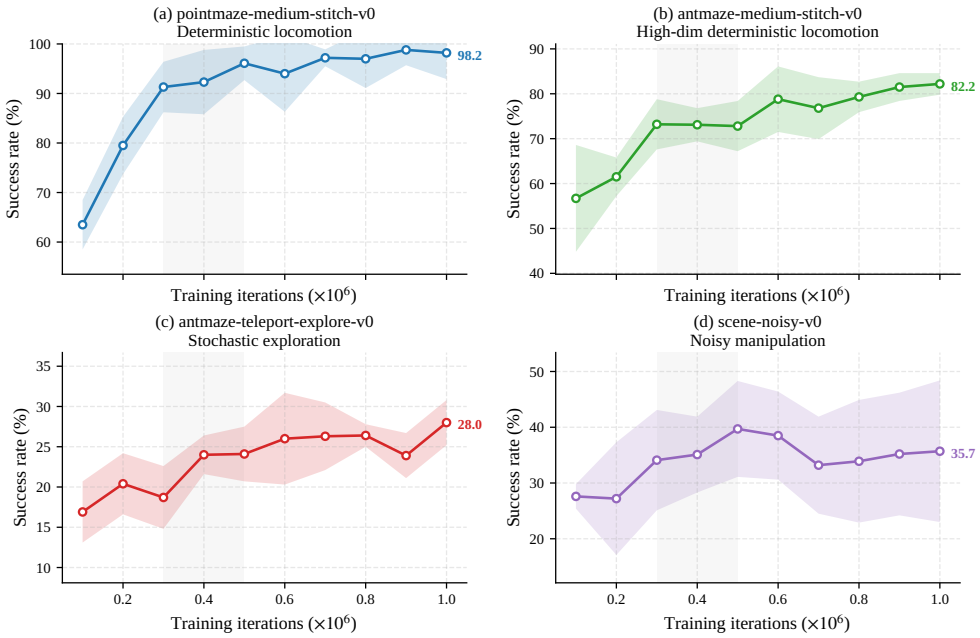
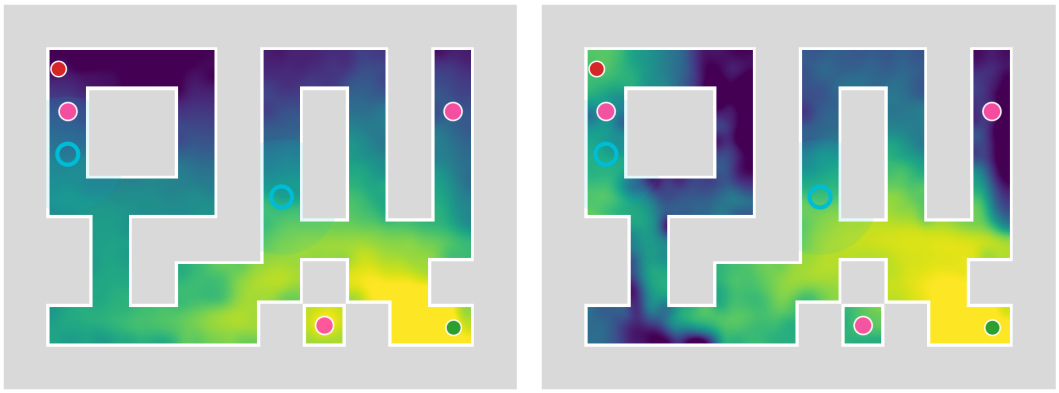


Figure 6: **Training curves of NFTR across four representative environments.** Solid curves show mean success rate over 8 random seeds. Shaded bands show ± 1 standard deviation. The light gray vertical band marks the 0.3 to 0.5×10^6 convergence window. The colored number on the right of each curve is the mean success rate at 1×10^6 steps. (a) Deterministic locomotion. (b) High-dimensional deterministic locomotion. (c) Stochastic locomotion, the regime in which Proposition 3.3(4) reduces to a one-sided bound and which therefore tests the practical stability claim most directly. (d) Noisy manipulation.

F.11 Visualization of Learned Representations

Both heatmaps in Figure 7 visualize the high-level subgoal preference for a fixed state–goal pair (s, g) (green/red dots). For each candidate waypoint w, we compute a target subgoal representation $z_{\text{tgt}} = \phi(s, w)$ and plot the relative density $\log p_\theta(z_{\text{tgt}} \mid s, g)$ over feasible maze regions (brighter $\Rightarrow$ higher preference within each panel). Gray blocks indicate walls / unreachable space. In the teleport maze, cyan hollow circles denote *teleport-in* portals (trigger locations), while magenta filled circles denote *teleport-out* landing sites.

In Figure 7 (a), HIQL's unimodal Gaussian produces a diffuse landscape that effectively averages across alternative routes, spreading preference beyond specific corridor structures. In contrast, Figure 7 (b) shows NFTR's Normalizing Flow yielding a more structured, multi-modal preference pattern, concentrating probability mass around distinct feasible regions, which better reflects multiple valid subgoal modes.



(a) HIQL (b) NFTR

Figure 7: Log-likelihood heatmap on antmaze-teleport-navigate `task1`.